November 2017

Using Exemplar Items to Define Performance Categories: A Comparison of Item Mapping Methods

Ana Karantonis

Using Exemplar Items to Define Performance Categories:

A Comparison of Item Mapping Methods

A Dissertation Presented

by

ANA KARANTONIS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements of the degree of

DOCTOR OF PHILOSOPHY

September 2017

College of Education

Using Exemplar Items to Define Performance Categories:

A Comparison of Item Mapping Methods

A Dissertation Presented

By

ANA KARANTONIS

Approved as to style and content by:

_____
Lisa A. Keller, Chair

_____
Ronald K. Hambleton, Member

_____
Craig S. Wells, Member

_____
Aline G. Sayer, Member

_____
Joseph B. Berger, Senior Associate Dean
College of Education

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor Lisa Keller for her unwavering support. Her guidance, friendship, strength, humor, patience and knowledge of all things statistical have been extremely valuable on this journey. I would also like to express my deep gratitude to the other members on my dissertation committee, Ron Hambleton, Craig Wells and Aline Sayer, for their guidance. In addition to their support during the dissertation process, their courses were outstanding and have proven extremely valuable throughout my career. In a similar light, I would also like to thank Steve Sireci for introducing me to a multifaceted validity framework which guides my work every day.

I would like to thank the Rhode Island Department of Education in general and MaryAnn Snyder and Phyllis Lynch in particular, for giving me the time and support to complete this dissertation. Furthermore, I would like to thank my colleagues Kate Schulz, Colleen O'Brian, and Patty Carnevale for their friendship and gentle nagging.

Finally I would like to thank three special boys in my life: Luke for giving me a reason to hope, Win for giving me a reason to fight, and John for being the best father to my sons.

ABSTRACT

USING EXEMPLAR ITEMS TO DEFINE PERFORMANCE CATEGORIES:

A COMPARISON OF ITEM MAPPING METHODS

SEPTEMBER 2017

ANA KARANTONIS, B.A., YALE UNIVERSITY

M.Ed., BOSTON COLLEGE

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by Professor Lisa A. Keller

Score reporting is an extremely important and yet often neglected component of large-scale assessment programs. One element of score reporting that frequently leads to misunderstanding is the interpretation of performance levels. One way to help define performance levels is through the use of "exemplars." Exemplars are test items that are supposed to best characterize each performance level. In this study, a Monte Carlo simulation was conducted to examine the performance of two item-mapping methods and different criteria for identifying exemplars under several simulated conditions.

The results of the study were neither clear nor systematic across all conditions and performance levels; however, there were a few findings. Using a discrimination criteria in addition to using RP alone, improved the false positive rate results for both tests. The converse was true, however, for the true positive rate results. Results showed that using a discrimination criterion in addition to using RP alone, decreased the true positive rates. With respect to both true positive and false positive rates, results under the normal distribution condition appeared better than under the skewed distribution condition for

the Empirical-based method but no clear patterns were observed between the two distributions for the Model-based method, suggesting that the Model-based method may be less susceptible to changes in the shape of the distribution than the Empirical-based method.

The study suggests that several factors should be considered when choosing item-mapping methodology for the purposes of identifying potential exemplars: number of exemplars desired, distribution of item difficulty across scale, shape of ability distribution, and resources available for content specialists to subsequently review the potential exemplars.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background

Score reporting is an extremely important and yet often neglected component of large-scale assessment programs. In fact, Goodman and Hambleton have argued that while "a great amount of attention has been directed toward the creation of technically sound assessments that can stand up to intense public and professional scrutiny[,] considerably less attention…has been given to ways in which the results of the assessments are organized, reported, and used" (2004, pp. 145-146). The most sophisticated measurement advances will be wasted if, at the end of the day, intended audiences do not understand what the reported test scores mean. How useful can test results be when students, parents, educators, and the public at large frequently misinterpret those results?

Measurement specialists have recognized the importance of score reporting and have called for research to improve score reporting practices for decades. In 1994, for example, Hambleton and Slater made a strong plea for improved score reporting practices:

> …without improvements to our scales and reporting forms, no matter how well we construct tests and analyze data, we run the serious risk of being ignored, misunderstood, or judged as irrelevant. The challenge to measurement specialists is clear. We now need to get on with the research. (p. 22)

A decade later, the need for improved score reporting practices was still evident:

> Very little research currently exists on how student-level results from large-scale kindergarten to Grade 12 assessments are reported. Given the increased role test results will play in the United States as a consequence of NCLB and the available evidence that shows the difficulties that many

> people have in understanding large-scale assessment results, there is a clear need to identify effective ways to report student-level results on large-scale assessment. (Goodman & Hambleton, 2004, p. 146)

As alluded to in the previous quotation, the enactment of the No Child Left Behind (NCLB) Act of 2001 shed a spotlight on test results, and by extension, score reporting practices.

NCLB brought renewed interest in score reporting in several ways. First, NCLB increased significantly the amount of K-12, large-scale testing being conducted in the U.S. by requiring that all states conduct annual testing in reading and mathematics in grades 3-8 and at least once in high school. States were also required to conduct annual testing in science at least once in grades 3-5, at least once in grades 6-9, and at least once in grades 9-12. This increase in the administration of large-scale assessments meant that more students, parents, and educators were looking at test score reports and trying to make sense of them. Second, NCLB made explicit requirements with regards to score reporting. Namely, NCLB required states to provide:

> individual student interpretive, descriptive, and diagnostic reports...that allow parents, teachers, and principals to understand and address specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic standards, and that are provided to parents, teachers, and principals, as soon as is practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable in a language that parents can understand. (NCLB, 2002, §1111[b][3][C][xii])

Finally, the high-stakes consequences associated with poor performance on NCLB-mandated tests focused the attention of educators more than ever on trying to increase test scores. Educators began to demand, therefore, that score reports provide information that will help improve instruction. As noted by Ryan (2003), Departments of Education began

to receive “increased numbers of requests for assessment information that could be used to review and guide instruction” (p. 1).

More recently, the enactment of the Every Student Succeeds Act (ESSA) of 2015, may have lessened some of the high stakes associated with large scale testing (for example, prohibiting the federal government from mandating educator evaluation systems) but it did nothing to reduce the number of grades and subjects tested under NCLB nor did it change the requirements with regards to score reporting (ESSA, 2015). As such, finding effective ways to report results on large-scale assessments continues to be a pressing need in the education measurement field.

1.2 Performance Level Interpretations (and Misinterpretations)

One element of score reporting that frequently leads to misunderstanding, and which is at the core of what educators are requesting, is the interpretation of performance levels. Under NCLB and ESSA, states must report results in terms of percentages of students in at least three performance levels. Specifically, states are required to develop challenging academic standards that

> …describe two levels of high achievement (proficient and advanced) that determine how well children are mastering the material in the State academic content standards; and…describe a third level of achievement (basic) to provide complete information about the progress of the lower-achieving children toward mastering the proficient and advanced levels of achievement. (NCLB, 2002, §1111[b][1][D][ii])

But what does it mean that a student is "proficient" in math? What does he/she know that a “non-proficient” student does not know? What can he/she do that the "non-proficient" student cannot do? As early as 1951, Flanagan realized that “test scores are meaningful and valuable to the extent that they can be interpreted in terms of capacities, abilities, and accomplishments of educational significance” (as cited in Ryan, 2003, p. 1).

It is imperative, then, that performance level descriptions, in order to be useful, include statements about or examples of the knowledge and skills students know and can do.

One of the reasons why performance levels are often misunderstood is that similar or identical labels (e.g., basic, proficient, advanced) are frequently used to describe performance levels on different assessments. Although the labels may be similar, their interpretation can be, and often are, quite different. Even within the context of NCLB, where all students were supposed to reach "proficiency" by the year 2014, each state was responsible for developing assessments that were aligned to its own state academic standards. Moreover, each state was responsible for defining what "proficiency" means with respect to those standards. Because performance levels have different meanings for different assessments, it is important to clearly define meanings for reported performance levels that are specific to particular assessments. How can this be done, however, in a way that is useful, meaningful, and easily understood by intended audiences?

1.3 Exemplars and their Applications

One way to help define performance levels is through the use of "exemplars." Exemplars are test items that are supposed to best characterize each performance level. Exemplars can be used in two ways to help define performance levels. First, they can be released to the public as examples of the types of items students performing at each level are likely to answer correctly. Second, exemplars can be used to help write performance level descriptions of the types of knowledge and skills students performing at each level are likely to know and be able to do.

Regardless of how exemplars are used, both applications typically rely on a statistical and a judgmental component. In both applications (either for the public release of sample items or for the writing of performance level descriptions), the statistical component involves identifying the items that *potentially* best describe each performance level. In the first application, once potential exemplars have been identified, a group of experts will typically use a judgmental process to select the *final* set of exemplars that will be released to the public. In the second application, once potential exemplars have been identified, a group of experts will typically use a judgmental process to write *descriptions* of what students at each performance level know and can do. While the judgmental components are essential to both processes and should be further researched, the present study will focus solely on the statistical component. Specifically, this study will address the following questions:

- How to identify potential exemplars using statistical methods?
- How do the available statistical methods perform relative to each other?

1.4 Item-mapping

Why is the statistical component important or even necessary? It is possible to convene a panel of content experts to review all the test items and item statistics on a particular test and have the panel select final exemplars for public release. It is also possible to convene a panel of content experts to review all the test items and item statistics on a particular test and have the panel write performance level descriptions based on those data. To do so, however, would require a lot of time and resources to first explain the statistical jargon to the panelists, and second, to review all the items on the test. If a subset of items (potential exemplars) can be identified ahead of time through

more efficient, statistical means, the cognitive load required of the panel of experts would decrease, and the feasibility of providing exemplar-based information to the public would increase considerably.

The statistical method for selecting potential exemplars is often referred to as item-mapping. As defined in Hambleton and Slater (1994), item-mapping refers to the process of locating an item on the test score scale.[1] Item-mapping has been used for purposes other than selecting exemplars in educational assessment. Examples of such purposes include standard setting (e.g., see Cizek, 2001 for examples of standard setting methods that make use of item-mapping), and helping to interpret scales (e.g., see Ryan, 2003 for examples of the use of item-mapping in defining Rasch scales).

Item-mapping typically relies on item response theory (IRT) to locate items on the score scale using a response probability (RP) criterion. For example, Figure 1.1 depicts two item characteristic curves (ICCs) estimated using a one-parameter logistic model (see Hambleton, Swaminathan, & Rogers (1991) for a description of IRT models). By finding the location on the test score scale where examinees have an 80% probability of answering the Item 1 correctly (RP80), Item 1 is "mapped" to a test score of 200. Similarly, using an RP80, Item 2 is mapped to a score of 300. Note that different RP values would result in different item mappings. How are these item mappings useful in helping interpret performance levels?

[1] The term item-mapping has been used in the literature to refer to a number of different processes. In this study, the term item-mapping will be used to refer to the process defined in Hambleton and Slater (1994) described above.

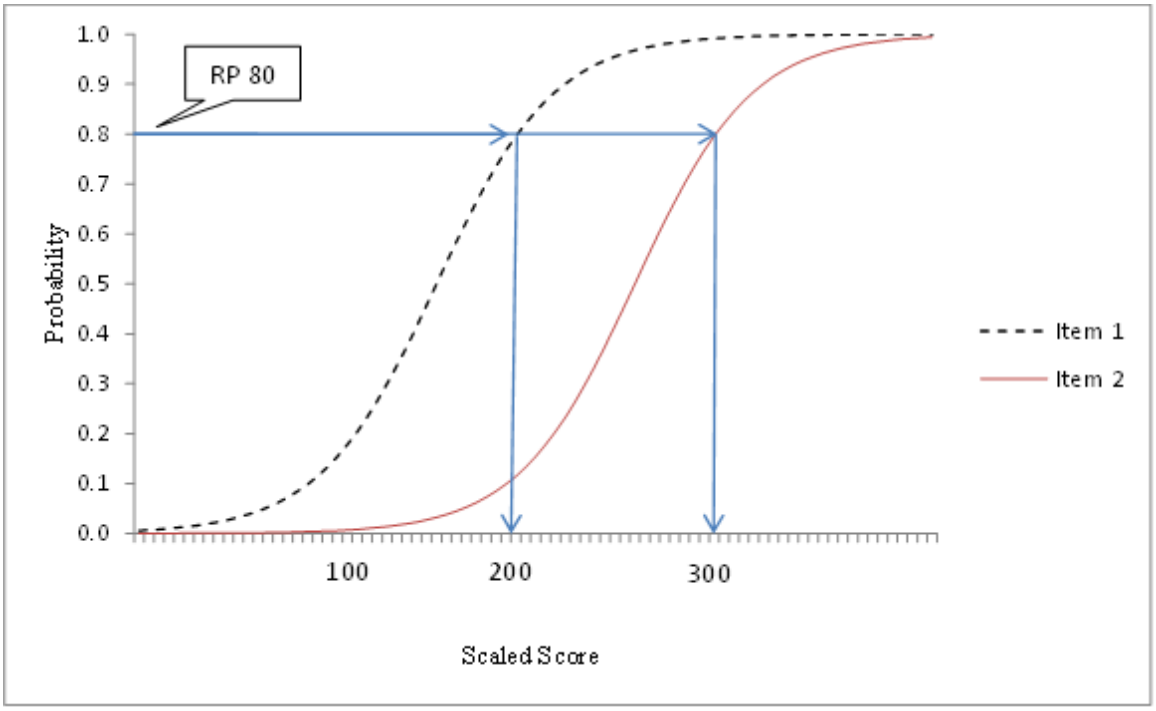

Figure 1.1 Illustration of item-mapping using item characteristic curves

Item 1 is an item that students with a score of 200 are likely (with an 80% probability) to get correct. By examining the knowledge and skills required to answer Item 1, and other items with similar mappings, content experts may be able to write statements about what students with scores around 200 are able to do. Similarly, by releasing Item 1 to the public, along with other items with similar mappings, educators may be able to determine the types of knowledge and skills needed to achieve a score of 200 on this particular test. While the above illustration described discrete score points on the scale (e.g., 200 and 300), the same methodology can be extended to include a *range* of score points (i.e., performance levels). Statements can then be made about the knowledge and skills examinees at particular performance levels will likely know and be able to do.

To aid in the interpretation of performance levels, item-mapping has been used in large-scale assessments such as the National Assessment of Educational Progress (NAEP) and the National Adult Literacy Survey (NALS) (Kolstad et al., 1998). The item-mapping method for selecting those items relied primarily on RP criteria. Using NAEP as an example, items must have an average percent correct of at least 50 for the particular achievement level to be considered as a potential exemplar item. "The common sense argument [for using the RP 50] is that we are justified in saying that students at a particular scale level 'can do' a task if the number of students who can do the task exceeds the number of students who cannot do the task" (Zwick, Senturk, Wang, & Loomis, 2001, p.16).

One problem of using an RP criterion to identify potential exemplar items is the lack of agreement regarding which RP value to use (e.g., see Karantonis & Sireci, 2006). Another problem is that some items may meet the RP criterion but do not discriminate well between adjacent performance levels. For example, suppose that for item *i* the average percent correct for students in the "proficient" level is just above the RP criterion (e.g., $p = 0.51$), and the average percent correct for students in the "below proficient" level is just below the RP criterion (e.g., $p = 0.49$). In this scenario, item *i* fails to discriminate between "proficient" and "below proficient" students since both groups tend to perform comparably.

To address these limitations, a *discrimination criterion* has also been used in conjunction with the RP criterion to help identify exemplar items (e.g., Beaton & Allen, 1992; Zwick et al., 2001). In 1992, Beaton & Allen described two item-mapping techniques used in NAEP that rely, in part, on discrimination criteria. The *direct method*

used item response data to calculate the proportion of correct responses at different anchor points on the scale. The *smoothing method* relied on monotonically increasing curves (e.g., three-parameter logistic curve) to estimate the proportion of correct responses at each point on the scale. With both methods, potential exemplar items were chosen by first imposing an RP criterion: only those items where a "substantial majority" of students correctly responded at particular anchor points were chosen. Both RPs of 80 and 65 had been used operationally in NAEP to define a "substantial majority" of students. Additionally, a discrimination criterion was imposed to identify potential exemplar items: only those items that presented a difference of 30% or higher in the proportion of correct response between adjacent anchor points were considered (Beaton & Allen, 1992).

In 2001, Zwick et al. extended this work by evaluating four methods for item mapping for use with NAEP (details of the study can be found in Chapter 2). The methods were evaluated based on the degree of consistency between method results and expert judgments regarding item difficulty, as well as on the ability of the methods to produce an adequate number of exemplar items. Additionally, the study examined the impact of imposing an extra discrimination criterion. The study concluded that it was preferable not to use a discrimination criterion because it resulted in fewer numbers of potential exemplar items and because it did not result in a higher degree of consistency with expert judgments (Zwick et al., 2001).

A limitation of the Zwick et al. (2001) study was that one of the criteria used to compare the methods was based on human judgments which contain error. Therefore, it would be interesting to compare the two methods when the “true” discriminatory power

of each item is known. In the present study, item-mapping methods that rely on RP criteria alone and a combination of RP and discrimination criteria will be examined through a Monte Carlo simulation study where the "true" RP and discrimination properties of the items will be known.

The literature available on item mapping for the purposes of identifying exemplars is limited. While a number of papers have described the process of item-mapping for other purposes (e.g., standard setting), only a few papers have discussed item-mapping within the context of exemplars selection. The majority of papers that do discuss item-mapping within the context of exemplars selection focused specifically on NAEP, where item-mapping methodology was first used to identify potential exemplars, or on NALS. Although NAEP and NALS are important testing programs in the U.S., these programs differ dramatically from most statewide testing programs; therefore, methodologies used in NAEP or NALS may not generalize to other programs. Thus, there is a need for extending the literature on item-mapping to other testing programs. Some of the concerns in NAEP, such as the small number of items available for public release, may not pose a problem to statewide testing programs where large number of items are typically released to the public after each test administration. The criterion used by Zwick (to maximize the number of items identified as potential exemplars) may not, therefore, be a priority for other testing programs. In the present study, data and test conditions will be simulated to mimic operational statewide assessments with the hope that the results from the study can generalize beyond assessments like NAEP and NALS.

No studies to date have explored the impact of variables such as sample size or shape of ability distributions on item-mapping results. Finally, no studies to date have employed a Monte Carlo simulation study to compare different item-mapping methods.

1.5 Purpose

The purpose of the present study is to add to the much needed body of research on item-mapping methods. Specifically, a Monte Carlo simulation study will be conducted to examine the performance of several item-mapping methods for identifying exemplars. A number of simulated conditions will allow for the examination of both empirical and model-based methods, the examination of different criteria for selecting items (RP alone or RP and discrimination combined), and the examination of other variables such as sample size and shape of ability distributions.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Overview

The present review of the literature on score reporting in general and on the interpretation of performance levels in particular has uncovered the need for future research to ensure that test scores are appropriately understood by intended audiences. This chapter will begin with a review of the 1999 and 2014 versions of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; AERA, APA, & NCME, 2014), followed by a summary of findings from recent literature reviews on score reporting. Examples of how performance levels in particular have led to misinterpretations will follow. Finally, research studies specifically related to the use of item-mapping methodology for the purposes of selecting exemplars will be reviewed.

2.2 Score Reporting

2.2.1 Standards

The importance of score reporting is evident by the emphasis placed on score reporting practices in both the 1999 and 2014 versions of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999; AERA, APA, & NCME, 2014). With their heavy emphasis on validity, the *Standards* provide explicit guidelines for the interpretation of test scores. As defined in the *Standards*, validity "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME , 1999, p. 9). As such, concerns about

whether score reporting practices aid or hinder the interpretation of test scores can be found throughout the *Standards*.

A thorough review by Ryan (2006), of sections of the *Standards* relating to score reporting found that several (9) of the standards explicitly required that test developers support valid test score interpretations through their reporting practices. Among them and particularly relevant to this paper are the following two standards:

> **Standard 5.10**
> When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of the scores, and how scores will be used. (AERA, APA, & NCME, 1999, p.65)

> **Standard 11.18**
> When test results are released to the public or to policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretation of the data. (AERA, APA, & NCME, 1999, p.117)

These two standards stress the responsibility on the part of test developers not only to provide clear statements about the meaning of test scores but also to anticipate any possible misinterpretations.

Additionally, Ryan found three standards that address group-level reporting. One of these standards, found below, is the only standard that addresses performance level reporting and is, therefore, particularly relevant to this paper.

> **Standard 8.8**
> When score reporting includes assigning individuals to categories, the categories should be chosen carefully and described precisely. The least stigmatizing labels, consistent with accurate representation, should always be assigned. (AERA, APA, & NCME, 1999, p.88)

In addition to requiring careful consideration when choosing labels to define the performance levels, Standard 8.8 also stresses the importance of describing "precisely" what each category means. Unfortunately, the *Standards* give no guidance regarding best practices for developing those performance level descriptions.

Finally, Ryan found two standards that applied to the timeliness of score reporting and one standard that applied to gain scores reporting, neither of which are relevant to this study.

The 2014 version of the *Standards* continue to place emphasis on the importance of valid score interpretations and score reporting. Standard 11.8 cited above, is still found in the new *Standards* in its original language under Standard 9.8 (pg.144) and Standard 5.10, now Standard 6.10 is also found in the new *Standards* albeit with minor edits.

> **Standard 6.10**
> When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (AERA, APA, & NCME, 2014, p.119)

The new *Standards* go further in recommending that test developers of educational tests provide "information that is understandable and useful to stakeholders" (pg. 194).To do so, test developers should engage in ongoing research (such as usability research with consumers) to improve the design and usefulness of score reports.

While the importance of score reporting practices in support of valid test score interpretations is evident in both versions of the *Standards*, the question remains whether research is being conducted to evaluate current score reporting practices and to develop

better reporting practices in the future. To answer this question, the following section will summarize the findings from reviews of literature on score reporting.

2.2.2 Reviews of Literature on Score Reporting

In 2004, Goodman and Hambleton examined score reports from 11 states, two Canadian provinces, and three U.S. testing companies in order to identify promising and problematic features included in score reports and to provide recommendation for future practices. As part this study, Goodman and Hambleton conducted a review of the literature on score reporting and concluded that "very little research currently exists on how student-level results from large-scale kindergarten to Grade 12 assessments are reported" (p. 146). Results from the literature review will be described followed by several recommendations derived from the study.

Goodman and Hambleton found that much of the research available on score reporting had been conducted within the context of the National Assessment for Educational Progress (NAEP). Concerns regarding some of the earlier NAEP score reports were described in this study and are listed below:

- [Reports] assumed an inappropriately high level of statistical knowledge for even well-educated audiences.
- Too many technical terms, symbols, and concepts were required to understand the message underlying even simple data.
- Statistical jargon (e.g., statistical significance, variance, standard error) confused and even intimidated some users.
- Symbols (e.g., "<" and ">" to denote statistical significant differences) and technical footnotes were misunderstood or ignored by many users of the reports.
- [Reports] presented too much information, making it difficult for readers to find and extract what they really want to know.
- [Reports included] overly dense displays that readers find daunting.
- [Reports did not make] enough use of graphical alternatives to textual and tabular forms.

- When attempts were made to redesign the displays for easy access (e.g., using three-dimensional bar and pie charts), they sometimes led to problems as increased clutter or perceptual inaccuracies.
- [Reports lacked] descriptive information (e.g., definitions and concrete examples) that would have helped provide meaning to the assessment results. (Goodman & Hambleton, 2004, p. 149)

Furthermore, the literature review uncovered several general principles for reporting results from large-scale assessments in an effective manner. These principles, listed below, were gleaned from literature on score reporting as well as from literature on the visual display of quantitative information.

(a) making the report readable, concise, and visually attractive;
(b) keeping the presentation clear, concise, and uncluttered;
(c) not trying to do too much with a data display (i.e., displays should be designed to satisfy a small number of preestablished purposes);
(d) including text to support and improve the interpretation of charts and tables;
(e) minimizing the use of statistical jargon;
(f) including a glossary of key terms;
(g) using bar charts to facilitate comparisons;
(h) grouping data in meaningful ways;
(i) using boxes or graphics to highlight main findings;
(j) avoiding the use of decimals;
(k) using color in a purposeful manner (given the potential for misuse, however, the general use of color was not universally recommended);
(l) piloting the reports with members of the intended audience;
(m) creating specially designed reports for different audiences. (Goodman & Hambleton, 2004, p. 150)

In their study, Goodman and Hambleton (2004) identified the strengths and weaknesses of score reports and interpretive guides from 11 states, two Canadian provinces, and three U.S. testing companies. From this analysis, the authors developed a set of recommendations for reporting student-level assessment results. These recommendations are as follows:

1. Include all information essential to proper interpretation of assessment results in student score reports (e.g., statements

explaining the purpose of the assessment, the meaning of performance levels and test scores, and how the test results should be used, and examples of how to interpret the confidence bands).
2. Include detailed information about the assessment and score results in a separate interpretive guide, ideally one in which the student score report can be inserted.
3. Personalize the student score reports and interpretive guides.
4. Include an easy-to-read narrative summary of the student's results at the beginning of the student report.
5. Identify some things parents can do to help their children improve. Ideally, these suggestions would be inserted in a separate section near the end of the score report and would be tailored to the student's performance. Advise parents and guardians to talk with their child's teacher about other ways to improve performance. (Goodman & Hambleton, 2004, p. 219)

In 2006, Ryan also conducted a review of the research on score reporting. Having found only a small number of more recent studies, the review relied primarily on the findings from Goodman and Hambleton (2004) described above. The author concluded from the review of the literature that "many educators have difficulty interpreting score reports from large scale assessment programs" (p. 705). Nevertheless, the author argued there are several features that can be manipulated to make score reports more informative as well as user-friendly. These features were characterized into two broader categories: *basic content,* and *format, language and display features*. Regarding *basic content*, the author recommended that "score reports should be related as closely and explicitly as possible to the content standards the assessment is designed to examine" (p. 705). Regarding *format, language and display features,* the author reiterated the principles provided in Goodman and Hambleton (2004), and strongly recommended the use of focus groups to evaluate the various aspects of score reports.

More recently, Hambleton and Zenisky (2013; Zenisky & Hambleton, 2012) reviewed the literature on score reporting and proposed a model for score report

development derived from that review and general best psychometric practices. The main contribution of this model was to formalize and standardize this important aspect of test development that has often been treated as an adhoc activity by many testing programs. The model is defined by seven steps and associated guiding questions:

1) Carry out needs assessment: What are the information needs of key stakeholders used to guide score report development?
2) Identify the intended audience(s): Who are the audiences for the score report and what audience characteristics should be considered to support the choice of information and level of detail needed?
3) Review report examples / literature: What does the literature contain regarding examples of student and parent reports or whichever reports are of interest?
4) Develop reports: In developing score reports, how can information from Steps 1, 2, and 3 be integrated into the process, and how are diverse talents involved?
5) Data collection / field test: How are reports field-tested?
6) Revise and redesign: How are the results from the field-test used in the redesign of the reports?
7) Ongoing maintenance: What is the plan to evaluate the reaction to the score report or reports when they are used operationally so that more revisions can be made for the next operational use?

The first three steps in the model can be taken simultaneously with the goal of clearly defining the purpose of the score report and gathering sufficient information needed to begin the active report development in Step 4. For Step 4, the authors strongly

recommend bringing in diverse talent (e.g., psychometricians, graphic designers, policy-makers, curriculum specialists, public relation specialists) to create draft reports. Additionally, the authors provide a very useful checklist which can be used during this design stage of the process. The checklist is divided into eight report element areas: 1) Needs assessment; 2) Content - Report introduction and description; 3) Content - Scores and performance levels; 4) Content – Other performance indicators; 5) Content – Other; 6) Language; 7) Design; and 8) Interpretive guides and ancillary materials. Once draft score reports are completed, the next step (Step 5) in the model is to field test those score reports. Field testing is a key component in test development but not often conducted for score reporting purposes. Step 6 (revise and redesign) should be seen as an iterative process, whereby the results of the field test are filtered through the design team and new drafts are created and field tested again as needed. The last step in the model is to develop a plan of ongoing maintenance whereby the use and utility of the score reports are continuously monitored.

2.3 Performance Levels Interpretations (and Misinterpretations)

The use of performance levels (also commonly referred to as performance standards, performance categories, or achievement levels) for the reporting of individual and group-level test scores had become popular even before the enactment of NCLB, which mandated that all states report performance-level scores. Their popularity was not surprising, as performance-level reporting provided criterion-based interpretations valued by the public (Koretz, 1995). In particular, assessment programs had increasingly begun to "report the percentage of students reaching or exceeding judgmental standards" (p. 284). In 1994, Hambleton and Slater similarly found that "performance standards were

greatly valued by policy-makers and educators" because they provided a "useful frame-of-reference for interpreting test score data" (p. 11).

While popular, the use of performance levels did not result in a panacea for test score reporting and test score interpretations. In fact, the use of performance levels, some have argued, has resulted in numerous misinterpretations on the part of the public. When NAEP began reporting scores using achievement levels labeled Basic, Proficient, and Advanced in 1990, for example, a slew of criticisms ensued (Koretz, 1995).

One controversial aspect of using achievement levels to report NAEP scores had to do with how the achievement levels were determined, or in other words, how the cut scores were set. While outside the scope of this paper, for further discussion on this matter, see National Academy of Education (1993) and Hambleton et al. (2000). For the present study, it will be assumed that performance levels have been established using appropriate and defensible methodologies and following guidelines such as those provided in Hambleton (2001).

Another aspect of using achievement levels in NAEP reporting that drew criticism had to do with how the achievement levels were described in the NAEP reports and the resulting misinterpretations by the public. In 1993, Koretz and Diebert published a study that examined the effectiveness of reports of NAEP achievement-level results. Effectiveness was measured through an analysis of articles published by the print media in the autumn of 1991, following the release of NAEP reports emphasizing achievement-level results. In reviewing the articles to determine the adequacy of the press accounts, the study found that "press accounts were often inadequate or even simply wrong" (p. 24). A number of key findings are noteworthy.

- Many of the articles reviewed showed no "understanding of the continuity of performance or clear notion of how to use […] achievement levels to place students on a continuum" (p. 24).
- Although the NAEP reports contained extensive descriptions of each achievement level, the media included only "sparse descriptions" of achievement levels. Often, no more than a one-word label ("Basic," "Proficient," "Advanced") or a two-to-three word description was provided ("solid academic performance").
- Where exemplar items were reported, it was found that the percentage of students reaching achievement levels was frequently confused with item p-values. Often these two concepts were used (incorrectly) interchangeably.

Based on the findings from this study, Koretz and Diebert (1993) argued that better methods were necessary to report NAEP achievement-level results. Particularly, NAEP score reports should include the following:

- Clear differentiation between actual and expected performance.
- Clearer ways of presenting actual performance on test items used to exemplify the reporting metric. Simply displaying seemingly inconsistent p-values along with the percentages of students reaching various levels on the scale has proven entirely insufficient.
- Explicit and concrete presentation of the continuity of student performance.
- Clear explanation of the role judgment plays in setting standards used for reporting and of the implications of the judgmental nature for proper interpretation of those levels.
- Clear and empirically defensible statements about what students at each of the reporting levels can do on the test. (pp. xii)

Also in an attempt to determine whether the public could understand NAEP's standards-based reports, Hambleton and Slater (1994) conducted a study examining

whether one of the NAEP executive summaries, *Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States*, was understood by its intended audience. This was accomplished by conducting a series of interviews with policy-makers, educators, and media representatives from a variety of states. Similar to Koretz and Diebert (1993), this study uncovered a number of elements in the NAEP reporting that were confusing to its intended audience. Some of the elements interviewees found confusing included:

- Statistical jargon such as descriptions of statistical significance and the symbols denoting statistical significance.
- The concept of "average proficiency scores" as determined by the NAEP scale was often confused with the category of "proficient students."
- The "standard errors" associated with and presented next to each percentage on a table.
- The percentage of students scoring *at or above* a proficiency level was often confused with the percentage of students in *each* proficiency level.
- Unfamiliar and complex chart formats.

The authors concluded that although standards-based reporting, "in principle," provides valuable information, there is a "need to focus considerable attention on the way in which scores are reported to minimize confusion as well as misinterpretations" (p. 21).

In 2003, Ryan examined six methods of reporting performance-level information to determine which were more easily understood by and valuable to its intended audience. Through a focus group consisting of 21 participants including teachers, principals, district curriculum and research/assessment directors, and state

curriculum/assessment specialists from South Carolina, the study examined the following six methods:

1. Item Content Objective Mapping – Graphical mapping of the content objectives associated with each item from a test form, multiple test forms, or the item bank on an ability/item difficulty scale with achievement-level cut scores reported.
2. Achievement Performance Level Narrative – Description of the content objectives assessed by the items at the various achievement levels, e.g., Below Basic, Basic, Proficient, Advanced.
3. Strand Achievement Levels for Individual Students – Mapping achievement-level cut scores from the total level to subscales or strand/areas for individual students and reporting by achievement level.
4. Strand Achievement Levels for Groups – Mapping achievement-level cut scores from the total test level to subscales or strands/areas for groups such as schools or school districts and reporting by achievement level.
5. Observed, Expected, and Differences in Strand and Item Performance for a Group – Observed strand/area and item performance (proportion answering correctly) for schools or districts relative to the proportion expected to answer correctly based on the groups' mean performance on the total test.
6. Observed, Expected, and Differences in Strand and Item Performance at the Achievement Level Cut Scores – Observed strand and item performance (proportions answering correctly) for schools or districts in comparison to the proportion of students in the state who are expected to answer correctly at each achievement-level cut score. (p. 34)

Results from the study showed that the two methods found to be "most useful" by the participants involved the use of exemplars and item-mapping. The highest rated method in terms of usefulness was Method 2, which involved verbal descriptions of the achievement levels based on a review of exemplar items. Participants liked this method because it contained only verbal descriptions (no tables, charts, or graphs), and the descriptions referenced content at a fairly fine level of detail. The main suggestion for improving this method involved presenting the verbal descriptions in a bulleted format

instead of paragraph form in order to "make the information clearer and easier to understand" (p. 50).

The second highest rated method in terms of usefulness was Method 1. In this method, a short description of the content objective of each item (e.g., prime factors, type of angle, range of data) was mapped onto the proficiency scale along with the performance-level cut scores. Based on focus group results, the author of the study characterized this method as one that "shows promise, has potential, but needs work" (p. 41). The following three shortcomings of the method were identified:

1. The item map was viewed as being too hard for many audiences to understand. There would be a need to provide additional interpretive materials and professional development to ensure the maps were being interpreted correctly.
2. The fact that the item-map presented to the study participants was based on a particular test form and that the item map would, therefore, change every year was seen as a drawback.
3. There was a concern regarding which response probability should be used to map the items. In this study an RP of 50 was employed. Some participants suggested that the RP should be higher. Participants also expressed their concern that regardless of what RP value to use, RP would be a difficult concept to explain to parents and teachers.

2.4 Exemplar Items and Their Applications

Exemplar items have been used in a number of ways in an attempt to make test scores more meaningful. In this section, the different ways in which exemplar items have

been used in NAEP will be described followed by the applications of exemplars to other assessment programs.

In 1985, an item-mapping technique was developed to report NAEP scores. "For each item on the assessment, the point on the scale was identified at which individuals with that level of proficiency had an 80 percent probability of responding correctly" (Phillips et al., 1993). A selection of items was then identified, paraphrased (into short descriptions of the item, e,g., "add two 3-digit number"), and displayed graphically along side the NAEP scale. This graphic display was referred to as an item map. Additionally, subsets of items were presented in their entirety with data that represented the "percentages of students performing at or above various levels on the scale" (p. 19). These exemplars were used in conjunction with the item maps to help give meaning to the NAEP scale.

A different technique, scale anchoring, was also implemented in NAEP to aid in interpreting the assessment results (Phillips et al., 1993). In this procedure, four anchor points were first arbitrarily chosen to partition the scale into five sections, or levels of proficiency. To give meaning to these anchor points, sets of exemplars, referred to as anchor items, were identified. These sets of exemplars were reviewed by panels of experts and used to articulate "the types of knowledge, skills, and reasoning abilities that were demonstrated by correct responses to the items in each set" (p. 27). The anchor item sets were identified through an item-mapping process.

When achievement levels for each grade level were introduced in the 1990 and 1992 mathematics NAEP assessments, exemplars were also identified and included in score reports (Phillips et al., 1993). However, in the 1992 implementation of the

achievement levels, potential exemplars were not first identified using an item-mapping procedure. Instead, a panel of educators selected the exemplar items "to exemplify the full range of performance of the intervals between the levels" (p. 38). The resulting exemplars were heavily criticized. A number of analyses conducted by Burstein et al. (1996), concluded that the "the exemplars as a set did not accurately characterize the performance of the groups in question" (p. 42). Subsequently, the process for identifying potential exemplars reverted back to item-mapping so as to be based on actual student performance (Koretz, 1995).

Outside the context of NAEP, there have been few mentions of the use of exemplars to help interpret test scores. In one study, the use of exemplars was mentioned within the context of the National Adult Literacy Survey (NALS). Similar to NAEP, in NALS, item maps are presented to "give more meaning to the reported scores" (Kolstad et al., 1998, p. 20). Additionally, the sets of items mapped on to the NALS scale are then used to "infer descriptions of the cognitive requirements of tasks at the various levels" (p. 20).

In 2007, ETS conducted a scale-anchoring study of the new TOEFL iBT reading test (Garcia Gomez, Noah, Schedl,Wright & Yolkut, 2007). In this study, exemplar items were used to write performance descriptors to help test takers interpret their performance. In this study, items on a particular form were divided into four categories: items that mapped onto each of the three performance levels plus items that did not map onto either of the performance levels. Then content experts were convened to "articulate the knowledge, skills, and abilities that were demonstrated by correct responses to the questions at each level" (p. 422). The resulting performance descriptors are now

displayed on TOEFL iBT score report but the authors of the study acknowledge several outstanding questions with the use of this methodology, namely: were the criteria for selecting exemplars too rigorous or not rigorous enough? Will the descriptors generalize across forms? And ultimately are the descriptors useful to test takers?

In 2008, a similar study was conducted for the College Board to "obtain clear, meaningful and instructionally relevant descriptions of ordered performance categories on the SAT" (Hambleton, Sireci & Huff, 2008, pg. 3). In this study panelists were convened to write performance category descriptions that would clearly delineate the various score ranges on the SAT scale (200 to 290, 300 to 390, 400 to 490, 500 to 590, 600 to 690, and 700 to 800). To do so, the types of problems or questions that students within each interval successfully answered had to be identified. An item mapping method was used to select and organize these exemplar items for panelists to review. As with TOEFL iBT study, the results and methodology from the SAT study were used to develop descriptors which are now routinely displayed in student score reports via the SAT Skills Insight website (Patelis & Matos-Elefonte (2009). Nevertheless, a few questions remain. The choice or RP value in the item mapping procedure (RP65), as acknowledge by the authors, was reasoned but still arbitrary (Hambleton, Sireci & Huff, 2008). Furthermore, panelists had to review over 300 items per content area which was extremely time consuming. Clearly more research is needed regarding both the choice of RP value and ways to reduce the subset of exemplars to a more manageable number.

Finally, in Goodman and Hambleton (2004), described earlier, only one of the 15 assessments examined in the study used illustrative items (i.e., exemplars) to supplement

performance-level descriptions. It is unknown the degree to which other state-wide assessments use exemplars in an attempt to make test scores more meaningful.

2.5 Item-mapping Methods for Identifying Exemplars

As described earlier, the preferred method for identifying potential exemplars in NAEP has been through the use of item-mapping. Item-mapping has been described in a number of NAEP reports and studies (e.g., Beaton & Allen, 1992; Beaton & Johnson, 1992; Mullis et al., 1990). This section will first describe the item-mapping methods used in NAEP and then review a study that compared the performance of different item-mapping methods.

In 1992, Beaton and Allen presented detailed descriptions of the two item-mapping methods used in NAEP for the purpose of scale anchoring. The purpose of scale anchoring, as has been described previously, is to give meaning to selected points on a scale. The basic idea as described by the authors is as follows: "to find out what students at points on the scale know and can do, one may look to see what students in the assessed sample who are estimated to have scores at or near those points know and can do, as evidenced by their item responses" (p.192). The authors caution, however, that although scale anchoring procedures may be applied to any ordinal scale, there is no guarantee that the procedures will result in useful descriptions of the anchor points. Poorly constructed tests may result in scale levels that may not be anchorable or with an insufficient number of anchor items to allow for meaningful descriptions. This is an important point raised first by Forsyth (1991), who argued that the extent to which anchor descriptions provide valid interpretations is determined by how well the content domains for the assessments are defined.

The first item-mapping method described by Beaton and Allen (1992) is the direct method. In the direct method, item response data are used to calculate the proportion of correct responses at different anchor points on the scale. The steps for the direct method follow:

1. Form K groups of examinees, G_k, such that all members of the kth group have scores x_i at or near the anchor point x_k'.
2. For each item, determine the proportion of students at or near the various anchor points who were able to answer the item correctly.
3. For the first anchor point, determine which items, if any, a substantial majority of students at that level was able to answer correctly.
4. For the second and succeeding anchor points, determine which items, if any, a substantial majority of examinees at that level was able to answer correctly that most of the students at the next lower anchor point could not.
5. Given the sets of items [from steps 3 and 4], attempt to generalize to the types or levels of performance characterized by these items. (pp. 195-197)

The second item-mapping method used in NAEP is the smoothing method. The smoothing method relies on monotonically increasing curves to estimate the proportion of correct responses at each point on the scale. The steps for the smoothing method follow:

1. Choose a curve to represent the relationship between the item responses and the scale scores.
2. For each item, fit the item characteristic curve to the u_{ij} and x_j and locate the points, $x_j^{(p)}$, such that the proportion passing item *j* is *p*.
3. For the first anchor point, determine which items, if any, a substantial majority of students at that level was able to answer correctly.
4. For the second and succeeding anchor points, determine which items, if any, a substantial majority of examinees at that level was able to answer correctly that most of the students at the next lower anchor point could not.
5. Given the sets of items [from steps 3 and 4], attempt to generalize to the types or levels of performance characterized by these items. (pp. 201-203)

It is important to note that although only two methods were described, a number of terms have to be operationalized for each method. There are, therefore, many different ways of implementing each of the methods. For example, in the direct method, it is necessary to define what "near" the anchor point means. In NAEP, "near" was defined as within 12.5 scaled points from the anchor point, but one can imagine that many other values could have been chosen instead. The term "substantial majority of examinees" (i.e., RP value) also needs to be defined. Beaton and Allen presented both RP65 and RP80 as possible choices but gave no preference of one over the other.

The only study to date that has compared different item-mapping methods for the purpose of selecting exemplars was conducted by Zwick et al. in 2001. Using data from the multiple-choice section of a NAEP Physical Science subscale, this study evaluated four item-mapping methods with respect to the following criteria:

- Does the method produce a reasonable number of exemplar items for each achievement level?
- Does the method produce results that are consistent across random samples and across NAEP "plausible values"?
- Are the results of the method supported by expert judgment about the difficulty of the items and their appropriateness as exemplars? (Zwick et al., 2001, p. 24)

The four item mapping methods compared in the study differed in the way that the probabilities of correct response were calculated. The methods differed on two dimensions. First, the methods differed with respect to reliance on an IRT model. Two *model* methods relied on the three-parameter logistic model to calculate the probability of correct response. These *model* methods are similar to the smoothing method described in Beaton and Allen (1992). Two *empirical* methods did not rely on an IRT model and are similar to the direct method described in Beaton and Allen (1992). Second, the methods

differed with respect to whether the probabilities of correct responses were calculated using an entire interval or a single point. Two *interval* methods made use of the whole achievement level interval to calculate the probabilities of correct response, whereas, two *midpoint* methods made use of the midpoint of each interval for calculating the probabilities. By crossing the two dimensions, the resulting four item-mapping methods examined in the study were as follow:

1. Model interval method
2. Model midpoint method
3. Empirical interval method
4. Empirical midpoint method

In addition to examining the four item-mapping methods, the study also investigated the use of different RP-value criteria: RP-50, RP-65, and RP-74. Additionally, the study examined the effect of imposing a discrimination criterion.

The major findings from this study are summarized below.

- There were near identical results between the model interval and model midpoint methods. As pointed out by the authors, this is an important finding because the midpoint method is much simpler to implement.
- For the empirical methods, results were stable across half-samples and across plausible values.
- A smaller number of exemplars were identified when a discrimination criterion was imposed. Nevertheless, all item-mapping methods produced a reasonable number of exemplars (at least three per achievement level).

- The model-based methods produced results more consistent with expert judgment than the empirical methods.
- Regarding RP-value criteria, the study found that the use of RP-65 and RP-74 resulted in the identification of more exemplars than the use of RP-50. Although counterintuitive, the reason for this has to do with a selection rule. According to this rule, if an item mapped at a lower achievement level it could not map at higher levels. In other words, items were not allowed to map at more than one achievement level.

2.6 Choice of Response Probability Criterion in Identifying Exemplar Items

As alluded to in various sections of this paper, different RP criteria for identifying exemplars have been used in the past or suggested in the literature. The debate over choice of RP extends beyond exemplar selection to other applications of item-mapping, such as standard setting. For a discussion of the debate over RP value with regards to the Bookmark standard setting method, see Karantonis and Sireci (2006) and Mueller, Schneider and Eagan (2008). In this section different RP-value criteria will be presented along with rationales for and against their use, and with examples of the operational use of RP value.

2.6.1 RP 50

As described in Zwick et al. (2001), RP 50 has been supported on both "common-sense and theoretical grounds" (p.16). The common sense rationale described by Zwick et al. (2001) is that "we are justified in saying that students at a particular scale level 'can do' a task if the number of students who can do the task exceeds the number of students who cannot do the task" (p.16). Kolstad et al. (1998) also provided support for RP 50

based on a theoretical argument, "if we trust the assumptions of the IRT model […] then the RP50 will provide the item mapping that is most consistent with the substantive meaning of the scales" (p. 51).

Arguments against the use of RP 50 surround the notion of mastery (Kolstad et al., 1998): "if one is going to say that people with a particular score on an assessment can successfully perform a particular assessment task, one wants to be fairly sure that a substantial majority of them can do it" (p. 11).

RP 50 was used by ACT to set achievement levels for NAEP in 1992 (Zwick et al., 2001) and by ETS in 2007 to write performance descriptors for the TOEFL iBT (Garcia Gomez et al., 2007).

2.6.2 RP 65 & 67

RP 65 was used for scale anchoring in NAEP beginning in 1986 (Kolstad et al., 1998). RP 65 was also used in the 2008 SAT study conducted by Hambleton, Sireci and Huff. RP 67 is the RP commonly used in the Bookmark standard setting method (Karantonis & Sireci, 2006). Arguments for RP 65 and 67 have been made on similar grounds, namely that these values are consistent with the mastery notion (Kolstad et al., 1998). In addition, a technical argument has also been proposed for RP 67 in particular by Huynh (2006), based on maximizing the information of the correct response under several IRT models.

2.6.3 RP 80

RP 80 was used in NAEP for scale anchoring from 1983 until 1986 when a switch was made to RP 65 (Kolstad et al., 1998). In 1992, RP 80 was also used in NALS for all its item-mapping, including setting cut scores. A debate ensued following the release of

the 1992 NALS results regarding the use of the RP 80. Some argued that this criterion was too stringent and led to misinterpretations by the public (Kolstad et al., 1998).

The choice of RP clearly makes a difference, as the number and selection of exemplars will differ depending on which RP criterion is used (Zwick et al., 2001). The literature reviewed provided no clear recommendations for the use of one particular RP over another. More research is clearly needed in this area.

2.7 Use of Discrimination Criterion to Identify Exemplar Items

Where discrimination values have been used as a criterion for identifying exemplars, the discrimination value has been calculated in the same manner: the discrimination value represents the difference between the probability of correct response for a particular achievement level (or score point), and the next lowest level (or score point). For example, if 60% of examinees in the "basic" level answered item A correctly, and 20% of the examinees in the "below basic" category answered item A correctly, the discrimination value for the "basic" category for item A is .40. While the discrimination values are calculated similarly, three different discrimination criteria were found in the literature.

Beaton and Johnson (1992) described two discrimination criteria that had been used in NAEP to select anchor items:

> (a) That 80% of the students at one anchor point answer the item correctly and that less than 50% of the students at the next lower level do.
> (b) That 65% of the students at the higher level respond correctly, that less than 50% do at the next lower level, and that the difference in the percentage passing is at least 30%. (p. 171)

Lastly, Zwick et al. (2001) described the discrimination criterion used by ACT, Inc. (the contractor for setting the achieving levels in NAEP since 1992):

> To meet the ACT discrimination criterion at a particular achievement level, an item must have a discrimination value that is at or above the 60th percentile of the distribution (across items) of discrimination values at that achievement level. (p. 19)

Similar to RP value, the choice of discrimination criterion will have an effect on the selection of exemplars. No studies to date have compared results from using different discrimination criteria in identifying exemplar items.

The present review of the literature has uncovered the need for future research on performance-level reporting. While the use of exemplars has been used in the past to help interpret performance levels, there are still many questions left unanswered. In particular, there is very little research regarding the best methodologies for identifying exemplars. In this study several item-mapping methods for identifying exemplars will be examined.

CHAPTER 3

METHODOLOGY

3.1 Overview

A Monte Carlo simulation study was conducted to examine the performance of two item-mapping methods (empirical and model-based) for identifying exemplars under a variety of simulated conditions. The simulation study was designed to mimic, to the extent possible, two statewide assessments (each from a different state) used in making NCLB decisions. Specifically, operational data from a 2002 administration of a Grade 10 science assessment (Test A) and a 2006 administration of a Grade 10 mathematics assessment (Test B) were used to generate the simulated data in the study, and were also used as a guide for choosing the different simulation conditions.

3.2 Data

For Test A, item response data were simulated to represent responses to a 44-item grade 10 science test. The test consisted of 40 multiple-choice items and four polytomously scored (on a four-point scale), extended-response items.

For Test B, item response data were simulated to represent responses to a 42-item grade 10 mathematics test. The test consisted of 32 multiple-choice items, four dichotomously scored short-answer items, and six polytomously scored (on a four-point scale), extended-response items.

The data for both tests were simulated using WinGen (Han, 2006) under a variety of conditions described in the sections below. IRT item parameter estimates found in technical documentation from the operational statewide assessments mentioned above were used as generating item parameter values. Item response data for the multiple-

choice items were generated using the three-parameter logistic model (3PLM). Data for the short-answer items were generated using the two-parameter logistic model (2PLM). Lastly, data for the extended-response items were generated using the generalized partial credit model (GPCM).

For each test, four cut scores were imposed on the θ distributions, dividing the ability scales into five performance categories. The cut scores were chosen so that the percentage of students classified as belonging to each of the five performance categories were similar for both the operational data and the simulated data. This was achieved by finding the θ values associated with the percentiles that represented the cut scores during the operational administrations of the assessments. The goal was to identify cut scores that would be reasonable within the context of statewide assessments but not necessarily exact. As such the cut scores were placed at -1.5, -.05, 0.5, and 1.5 on the theta distribution. A total of 16 simulation conditions were used to generate item response data as displayed in Table 3.2. These 16 conditions reflect the two tests, differences in the ability distributions, and different sample sizes used to generate the data. Ten replications were performed under the 1,000, 2,000, and 5,000 sample conditions and one replication was performed under the 50,000 sample condition resulting in the generation of 124 datasets. Each of these simulated conditions is discussed in more detail below.

Table 3.2 Simulated Conditions

	Simulated condition	# of conditions
Test type	Test A Test B	2
Ability distribution	Normal Skewed	2
Sample size	1,000 2,000 5,000 50,000	4
Total number of simulated conditions		**16**

3.2.1 Ability distributions

As described earlier, the literature review found no studies that examined whether the shape of the ability distribution had any impact on the results from item mapping methods and yet the ability distributions of statewide assessments vary widely. To examine whether there are differences in the item mapping results across different ability distributions, two ability parameter distributions were used to generate item responses. Under the first condition, ability parameters were sampled from a standard normal distribution, $\theta \sim N(0,1)$. The normal distribution was chosen arbitrarily to serve as a comparison. Under the second condition, an attempt was made to match the underlying ability distribution of one of the operational tests. This was done by calculating the skewness of the raw score distribution for the operational administration of the test and generating theta scores from a beta distribution (alpha = 10, beta= 3.73) with similar skewness. The resulting theta distributions had a skewness of -.50 and kurtosis of 0.0. On the raw metric scale, the skewness resulted in an average mean shift in points earned from 26.6 to 30.3 for Test A (out of 56 total possible points) and an average mean shift from 38.9 to 43.2 for Test B (out of 60 total possible points).

3.2.2 Sample Sizes

Sample size was manipulated to examine how the stability of the estimates would impact the item-mapping results. Four different sample sizes were used to generate item response data. The first sample size of 50,000 was chosen to reflect the full sample of a typical statewide test administration. Only one replication of this condition was conducted. Additionally, smaller sample sizes of 1,000, 2,000, and 5,000 were also analyzed. These smaller sample sizes were chosen to represent possible field test samples as well as possible early samples in the event that a state may want to release exemplars in conjunction with the public release of score reports. Ten replications of each of these sample sizes were conducted.

3.3 Item-Mapping Methods

Item-mapping methodology was applied to the 124 generated datasets to identify potential exemplars. A total of twelve item-mapping methods were examined. These item-mapping methods varied with respect to the factors identified as important in the review of the literature described earlier. Specifically, the twelve item-mapping methods varied with respect to method type (empirical or model-based), criteria for selecting potential exemplars (RP-value alone or RP value and discrimination combined), and differences in the RP value chosen (RP50, RP65 or RP 80). The method types and the criteria for selecting exemplars are described in the following sections.

3.3.1 Method Type

Two types of item-mapping methodologies were examined in this study, the *empirical-based* and the *model-based* methods. The two methods differ in the way that the probability of correct response for each item in each performance level is calculated.

In the former, how students actually performed on the items is used and in the latter, the IRT model is used.

The *empirical-based* method was applied as follows: For each item *i* and each performance level *j*, an empirical response probability, p_{ij}, was calculated. The empirical response probability was calculated by first identifying those examinees whose ability estimates (θ values) fell within the performance level. Examinees were classified into each performance level based on their raw score and adjusted (rescaled) cutscores for each replication. The empirical response probability equals the proportion of correct response (conditional p-value) for those examinees. For example, for item 1, the proportion of examinees in the "basic" category who answered the item correctly was calculated. This proportion represents the empirical response probability p_{ij}.

The *model-based* method was applied as follows: For each item *i* and each performance level *j*, a model-based response probability, p_{ij}, was calculated. First, item parameters were estimated using the appropriate IRT model (3PLM for multiple-choice items, 2PLM for short-answer items, and GPCM for extended-response items). The item parameters were estimated using the software PARSCALE (Muraki & Bock, 2003) and rescaled using the Mean and Sigma method (see Hambleton, Swaminathan, & Rogers (1991) for a description of the Mean and Sigma method). Then, the probability of a correct response associated with the midpoint of each performance level was identified. For the first and last categories, θ values of negative and positive two were arbitrarily chosen as the lower and higher bounds of the categories, respectively. For example, in Figure 3.1, the response probability for performance level 4 of item *i*, labeled p_{i4} was

calculated by taking the midpoint of category 4 (1.0) and finding the response probability associated with that point (.86).

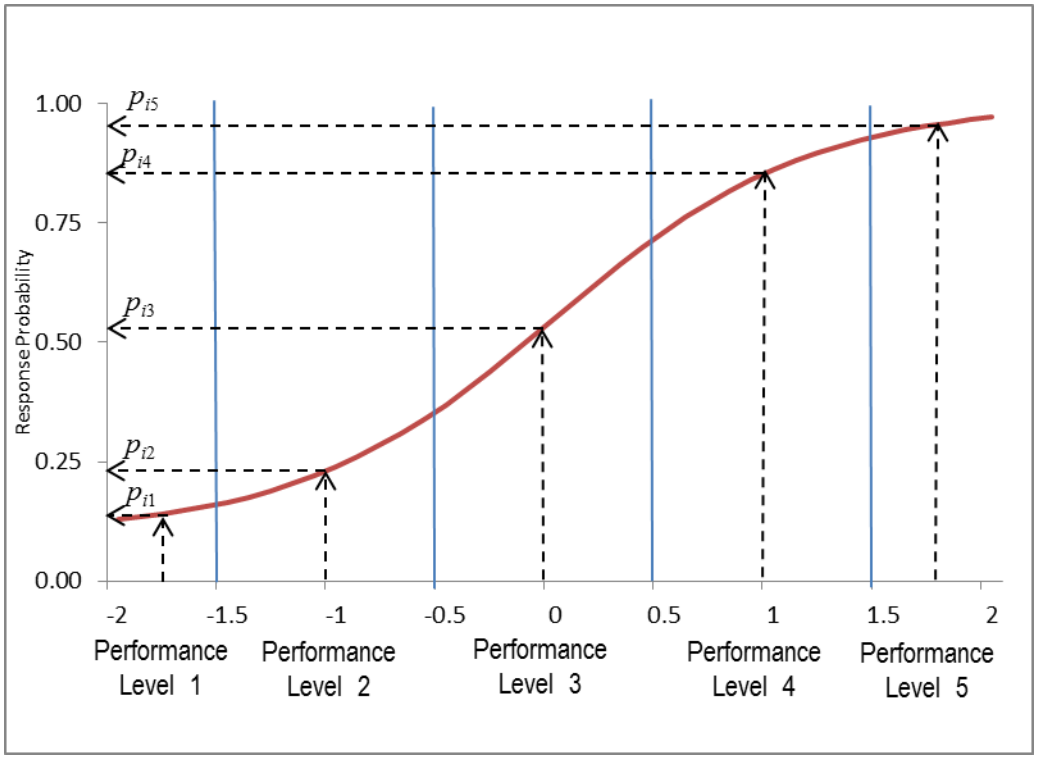

Figure 3.1 Illustration of Model-based Method

For each of these methods, the extended-response items were treated similarly. For the 4-point extended response items, each score point was viewed as a potential exemplar and was therefore treated as a separate item during the item-mapping process. As such, for both the empirical and model-based methods, the probability of correct response was calculated for each score point.

3.3.2 Criteria for Selecting Potential Exemplars

Once each item or score point was mapped using either the empirical or model-based methods, the items were selected using several criteria, described below. The goal was to select the items in the test that represented "potential exemplars" for each performance level except the lowest. For the lowest category, no exemplars were

identified because typically the lowest category is not associated with a description of skills (Zwick et al., 2001). Six criteria were examined to select the potential exemplars, which varied depending on whether RP-value alone or both RP-value and discrimination were used as criteria.

1) *RP 50*: The first set of items was selected based solely on an RP-value of 50. For each performance level (except the lowest), all items were identified whose response probability was greater than or equal to .50. Items were only allowed to map to one performance level. Therefore, items that were mapped to performance level 2 were no longer eligible to map to performance levels 3, 4, and 5. Similarly, items that mapped to performance level 3 were no longer eligible to map to performance levels 4 and 5. And items that mapped to performance level 4 were no longer eligible to map to performance level 5.
2) *RP 65*: The second set of items was selected based solely on an RP-value of 65. For each performance level (except the lowest), all items were identified whose response probability was greater than or equal to .65. As described above, items were only allowed to map to one performance level.
3) *RP 80*: The third set of items was selected based solely on an RP-value of 80. For each performance level (except the lowest), all items were identified whose response probability was greater than or equal to .80. As described above, items were only allowed to map to one performance level.
4) *RP 50 + Discrimination*: The fourth set of items was selected based on an RP-value of 50 and an additional discrimination criterion. For each performance level (except the lowest), all items were identified whose response probability

was greater than or equal to .50 AND at least greater than the previous performance level's response probability by .30. As described above, items were only allowed to map to one performance level.

5) *RP 65 + Discrimination*: The fourth set of items was selected based on an RP-value of 65 and an additional discrimination criterion. For each performance level (except the lowest), all items were identified whose response probability was greater than or equal to .65 AND at least greater than the previous performance level's response probability by .30. As described above, items were only allowed to map to one performance level.
6) *RP 80 + Discrimination*: The fourth set of items was selected based on an RP-value of 80 and an additional discrimination criterion. For each performance level (except the lowest), all items were identified whose response probability was greater than or equal to .80 AND at least greater than the previous performance level's response probability by .30. As described above, items were only allowed to map to one performance level.

3.4 Data Analyses

Results from this study were evaluated by comparing the identified potential exemplars obtained from all the replications against the "true" potential exemplars. The "true" exemplars were calculated by applying the model-based method to the generating item parameters instead of the item parameter estimates. The goal was to ascertain whether the "true" exemplars were identified more often than not. To do so, this study evaluated results both in terms of True Positive Rates (how often were the correct

exemplars identified) and False Positive Rates (how often were the incorrect exemplars identified).

CHAPTER 4

RESULTS

4.1 Overview

The results from this study are numerous due to the number of simulated conditions examined. To provide context for interpreting the results, first the IRT item parameter estimates from the two operational statewide assessments will be provided. These were the estimates assumed to be "true" and used to generate the simulated files. Additionally, the "true" exemplars under each item-mapping criterion (RP50, RP65, RP80, RP50 + discrimination, RP65+ discrimination, RP80+ discrimination) are identified. These data are presented to provide an overview of the relative difficulty and discrimination of all items along with the baseline of potential exemplars the Monte Carlo simulation will later attempt to recapture. Next, results from the Monte Carlo simulations for Test A will be presented followed by the results from Test B. Finally a summary of the results by item-mapping criterion will be presented at the conclusion of this chapter.

4.2 IRT Parameters and True Exemplars

Table 4.2.1 presents a summary of the IRT item parameter estimates for Test A. These were the estimates assumed to be "true" and used to generate all simulated files for Test A. The mean, minimum and maximum statistics are aggregated by item type, multiple-choice or extended response. The average difficulty (b-value) of the multiple-choice items was .377, ranging from -1.261 to 1.328. In terms of discrimination, the average a-value was .960, ranging from .348 to 2.122. Finally, the average guessing value (c-value) for the multiple-choice items was .164, ranging from 0 to .475. The extended-

response items for Test A were relatively more discriminating with an average a-value of 1.150. In terms of difficulty, the lowest threshold (d1) averaged -.789 while the highest threshold (d4) averaged 1.488.

Table 4.2.1 Summary IRT Parameters for Test A

Type	Statistic	A	B	C		
Multiple Choice	Mean	0.960	0.377	0.164		
	Minimum	0.348	-1.261	0.000		
	Maximum	2.122	1.328	0.475		
Type	**Statistic**	**A**	**D1**	**D2**	**D3**	**D4**
Extended Response	Mean	1.150	-0.789	0.017	0.859	1.488
	Minimum	0.975	-0.873	-0.043	0.592	1.125
	Maximum	1.287	-0.738	0.047	1.303	2.167

Table 4.2.2 presents the same information as Table 4.2.1 but in this instance the data presented refers to Test B. The average difficulty (b-value) of the multiple-choice items was -.822, ranging from -3.517 to .809. In terms of discrimination, the average a-value was .976, ranging from .356 to 1.690. The average guessing value (c-value) for the multiple-choice items was .145, ranging from 0 to .356. For the short-answer items, the average difficulty was -.533, ranging from -1.075 to -.045. The average discrimination for the short-answer items was .827, ranging from .588 to 1.070. Finally, the extended-response items, were once again relatively high discriminating with an average a-value of 1.405, and ranging from 1.160 to 1.823. In terms of difficulty, the lowest threshold (d1) averaged -1.590 while the highest threshold (d4) averaged .857.

Overall Test A appears to be more difficult than Test B. But both tests appear to be similar in terms of average discrimination and average guessing parameters.

Table 4.2.2 Summary IRT Parameters for Test B

Type	Statistic	A	B	C		
Multiple Choice	Mean	0.976	-0.822	0.145		
	Minimum	0.356	-3.517	0.000		
	Maximum	1.690	0.809	0.356		
Short Answer	Mean	0.827	-0.533			
	Minimum	0.558	-1.075			
	Maximum	1.070	-0.045			
Type	**Statistic**	**A**	**D1**	**D2**	**D3**	**D4**
Extended Response	Mean	1.405	-1.590	-0.630	0.071	0.857
	Minimum	1.160	-2.052	-1.077	-0.115	0.368
	Maximum	1.823	-0.963	-0.270	0.310	1.278

For item-level information, refer to Tables A.1and A.2, found in Appendix A, which provide the IRT parameters for each of the items in Tests A and B respectively.

Table 4.2.3 provides a summary of the number of "true" exemplars under each item-mapping criterion for Test A. These data are presented to provide the baseline of potential exemplars the Monte Carlo simulation will later attempt to recapture. Under the RP50 criterion, only two items were identified as true potential exemplars for Performance Level 2. Those two items represented 4% of all items on Test A. Both of those items were multiple-choice items and represented 5% of all multiple choice items on the test. No extended response item was identified as an exemplar for Level 2 under the RP50 criterion. Of note, only under the RP50 criterion were items identified as true exemplars for Level 2. Under all other item mapping criteria, no exemplars were identified. Another point to note is that the extended response score points were not identified as true exemplars for Levels 2, 3 and 4. Only for Level 5, were extended response score points identified as true exemplars.

Table 4.2.3 Number of True Exemplars for Test A

Item Mapping Criterion	Performance Level	All items		Multiple-Choice		Extended Response	
		#	%	#	%	#	%
RP50	Level 2	2	4%	2	5%	0	0%
	Level 3	12	21%	12	30%	0	0%
	Level 4	24	43%	24	60%	0	0%
	Level 5	5	9%	2	5%	3	19%
RP50 + Discrimination	Level 2	0	0%	0	0%	0	0%
	Level 3	2	4%	2	5%	0	0%
	Level 4	10	18%	10	25%	0	0%
	Level 5	4	7%	1	3%	3	19%
RP65	Level 2	0	0%	0	0%	0	0%
	Level 3	6	11%	6	15%	0	0%
	Level 4	24	43%	24	60%	0	0%
	Level 5	13	23%	10	25%	3	19%
RP65 + Discrimination	Level 2	0	0%	0	0%	0	0%
	Level 3	1	2%	1	3%	0	0%
	Level 4	11	20%	11	28%	0	0%
	Level 5	7	13%	4	10%	3	19%
RP80	Level 2	0	0%	0	0%	0	0%
	Level 3	2	4%	2	5%	0	0%
	Level 4	11	20%	11	28%	0	0%
	Level 5	26	46%	24	60%	2	13%
RP80 + Discrimination	Level 2	0	0%	0	0%	0	0%
	Level 3	1	2%	1	3%	0	0%
	Level 4	4	7%	4	10%	0	0%
	Level 5	6	11%	4	10%	2	13%

Similarly, Table 4.2.4 identifies the "true" exemplars under each item-mapping criterion for Test B. Once again these data are presented to provide the baseline of potential exemplars the Monte Carlo simulation will later attempt to recapture. In Test B, more items were identified as true exemplars than in Test A. For example, under the RP50 criterion, 17 items were identified as true potential exemplars for Performance Level 2. Those 17 items represented 28% of all items on Test B. Fourteen were multiple-choice items and represented 44% of all multiple choice items on the test. One was a short-answer item representing 25% of the short answer items on the test. And two were

extended-response score points representing 8% of extended-response score points on the test. For Test B, only under the RP80 and Discrimination criterion did Level 2 result in no exemplars identified.

Table 4.2.4 Number of True Exemplars for Test B

Item Mapping Criterion	Performance Level	All items		Multiple-Choice		Short Answer		Extended Response	
		#	%	#	%	#	%	#	%
RP50	Level 2	17	28%	14	44%	1	25%	2	8%
	Level 3	15	25%	12	42%	3	75%	0	4%
	Level 4	10	17%	6	17%	0	0%	4	16%
	Level 5	3	5%	0	0%	0	0%	3	12%
RP50 + Discrimination	Level 2	5	8%	4	13%	0	0%	1	4%
	Level 3	9	15%	6	25%	3	75%	0	4%
	Level 4	7	12%	4	11%	0	0%	3	12%
	Level 5	3	5%	0	0%	0	0%	3	12%
RP65	Level 2	10	17%	10	31%	0	0%	0	0%
	Level 3	13	22%	11	36%	2	50%	0	0%
	Level 4	16	27%	11	36%	2	50%	3	16%
	Level 5	3	5%	0	0%	0	0%	3	12%
RP65 + Discrimination	Level 2	2	3%	2	6%	0	0%	0	0%
	Level 3	7	12%	6	19%	1	25%	0	0%
	Level 4	11	18%	7	22%	1	25%	3	12%
	Level 5	3	5%	0	0%	0	0%	3	12%
RP80	Level 2	4	7%	4	13%	0	0%	0	0%
	Level 3	10	17%	10	28%	0	0%	0	0%
	Level 4	18	30%	13	47%	4	100%	1	8%
	Level 5	10	17%	5	14%	0	0%	5	20%
RP80 + Discrimination	Level 2	0	0%	0	0%	0	0%	0	0%
	Level 3	2	3%	2	6%	0	0%	0	0%
	Level 4	7	12%	5	17%	1	25%	1	4%
	Level 5	3	5%	0	0%	0	0%	3	12%

Of importance to note, for both tests A and B, the Response Probability and Discrimination criteria are by definition more strict than the Response Probability criterion alone. Therefore, the number of items identified as true exemplars with the discrimination criterion applied is always less than when RP is used as the sole criterion.

In the following sections, results from the Monte Carlo simulation studies will be presented. The questions to be answered are:

1) How often incorrect items are identified as exemplars (False Positive Rate results), and
2) How often the true exemplars are correctly identified (True Positive Rate results).

Figures 4.3.1 through 4.4.4 display the results for Test A and B under the different simulation conditions. The graphs on the left present the results for the Model-based method whereas the graphs on the right present the results for the Empirical-based method. Furthermore, the graphs are organized by sample size with the smallest sample size (1,000) presented first at the top of the page and the largest sample size (50,000) presented last at the bottom of the page. Within each graph, results for each item-mapping criterion (RP50, RP65, RP80, RP50 + discrimination, RP65 + discrimination, RP80 + discrimination) are represented as separate lines. The horizontal axis represents each of the Performance Levels with the exception of Level 1 for which, as described earlier, exemplars are not typically identified. The vertical axis represents either the False Positive Rate results (percentage of times “true” non-exemplar items were incorrectly identified as exemplars) or the True Positive Rate results (the percentage of times the "true" exemplars were correctly identified). Please note that for each of the figures 4.3.1 through 4.4.4, there is a corresponding table (tables B.1 through B.8) presented in Appendix B. Appendix B presents the same information as figures 4.3.1 through 4.4.4 but in tabular form and with the addition of an “N” size representing either the number of “true” non-exemplars for the False Positive Rate results tables or the number of “true” exemplars for the True Positive Rate results tables.

4.3 Test A results

Figure 4.3.1 displays the False Positive Rate results for Test A under the normal distribution condition. Beginning with the Model-based method under the 1,000 sample condition, the false positive rate is low for Performance Level 2 with all criteria showing 0% or 1% with the exception of RP50 which had a 6% false positive rate. Similarly, for Performance Level 3, the false positive rates were all at or below 4%. For Performance Level 4, the false positive rates were all at or below 5% with the exception of RP65 which had a 7% false positive rate. At Performance Level 5, the false positive rates were all at or below 2% with the exception of 7% for the RP80 criterion.

Under the 2,000 sample condition, the results show a similar pattern albeit the false positive rates for Performance Level 4 were generally smaller than under the 1,000 sample condition. For Performance Level 2, the false positive rates were all 0% with the exception of RP50 which showed a 6% false positive rate and RP50D which showed a 2% rate. For Performance Levels 3 and 4, the false positive rates were all at or below 3%. At Performance Level 5, the false positive rates were all at or below 5%.

Results for the 5,000 sample condition were nearly identical to the 2,000 sample condition. For Performance Level 2, the false positive rates were all 0% with the exception of RP50 which showed a 6% false positive rate and RP50D which showed a 2% rate. For Performance Levels 3, 4 and 5, the false positive rates were all at or below 3%.

And finally for the 50,000 sample condition the false positive rates were slightly lower across the board. For Performance Level 2, the false positive rates were all at or below 4%. For Performance Level 3, the false positive rates were all 0% with the

exception of RP50D with a 2% rate. Performance Level 4 showed the smallest false positive rates, all at 0%. And at Performance Level 5, the false positive rates were all 0% with the exception of RP65 with 2%.

The False Positive Rate results for the Empirical-based method were a bit different particularly for Performance Level 5. Under the 1,000 sample condition, the false positive rates for Performance Level 2 ranged between 0% and 6% (for RP50). For Performance Level 3 the false positive rates were low, all at or below 2%. For Performance Level 4, the false positive rates were at or below 5%. At Performance Level 5, however, the false positive rates ranged from 2% for the RP50 and RP50D criteria to 13% for RP80.

Under the 2,000 sample condition, the results show a similar pattern. For Performance Level 2, the false positive rates for Performance Level 2 ranged between 0% and 5% (for RP50). For Performance Levels 3 and 4, the false positive rates were all at or below 2%. Similar to the 1,000 sample condition results, at Performance Level 5, the false positive rates ranged from 3% to 14% (for RP80).

Results for the 5,000 sample condition were nearly identical to the 2,000 sample condition. For Performance Level 2, the false positive rates for Performance Level 2 ranged between 0% and 6% (for RP50). For Performance Levels 3 and 4, the false positive rates were all at or below 1%. At Performance Level 5, the false positive rates ranged from 1% to 12% (for RP80).

And finally for the 50,000 sample condition results improved slightly for Performance Levels 2, 3, and 4 but not for Performance Level 5. For Performance Level 2, the false positive rates were all at or below 4%. For Performance Level 3, the false

positive rates were all 0% with the exception of RP65 with a 2% rate. At Performance Level 5, however, the false positive rates for RP65 and RP80 continued to be relatively high with rates of 12% and 17% respectively. Of note, the false positive rates for Performance Level 5 did not appear to improve with the increase of sample size under the Empirical-based Model.

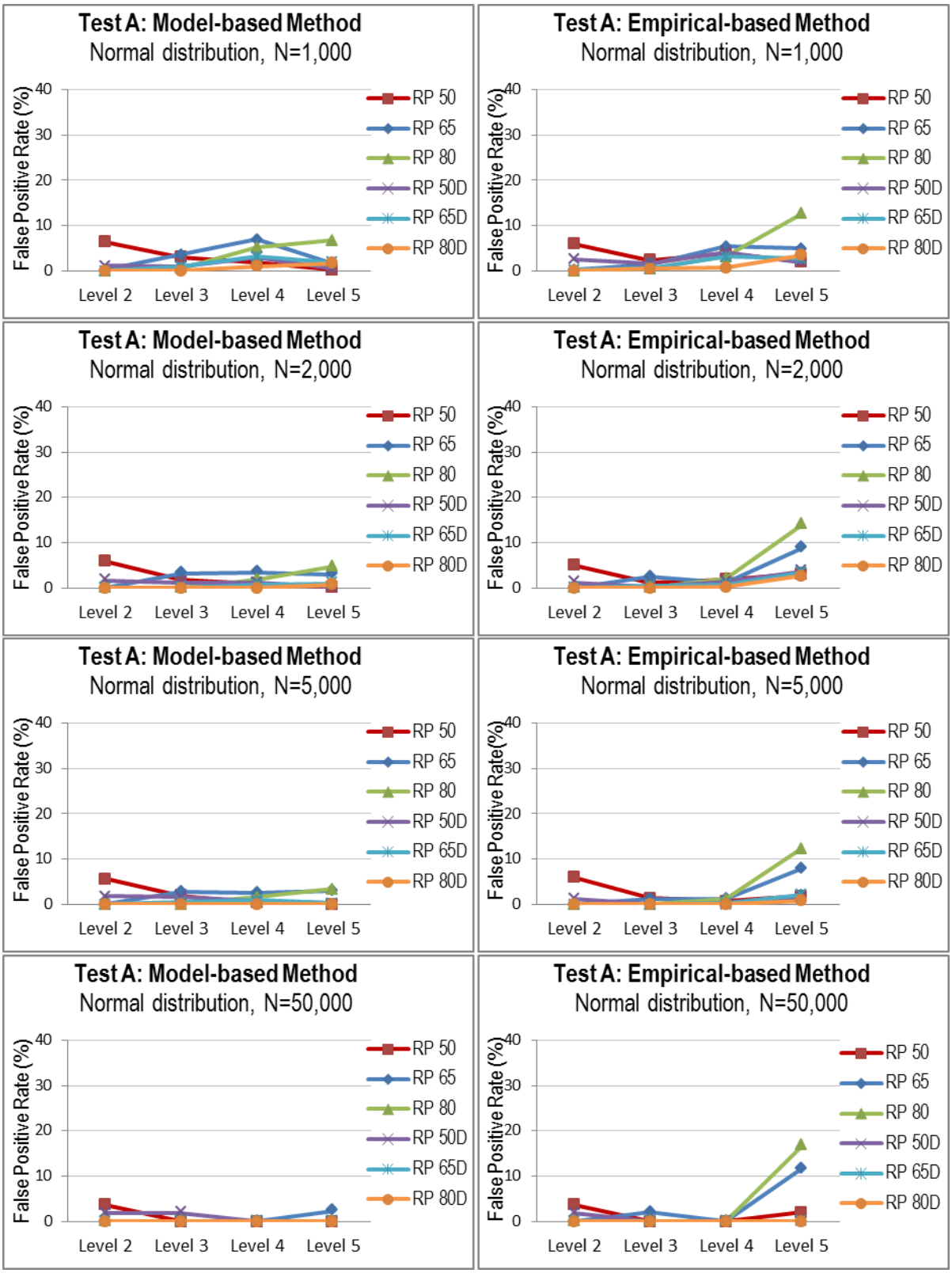

Figure 4.3.1 False Positive Results for Test A under Normal Distribution Condition

Figure 4.3.2 displays the False Positive Rate results for Test A under the skewed distribution condition. Similar to the results under the normal distribution for the Model-based method, under the skewed distribution condition, the false positive rate results were relatively low and improved with the increase in sample size. For example, for Performance Levels 2, 3, and 4, under the 1,000 sample condition, the false positive results were all at or below 5%. And for Performance Level 5, the only false positive rate above 3% was for the RP80 criterion with an 8% rate. With an increase in sample size, results improved slightly with all false positive rates at or below 5% for the 2,000 sample condition, at or below 4% for the 5,000 sample condition, and at or below 2% for the 50,000 sample condition.

The False Positive Rate results for Test A under the Empirical-based Method and skewed distribution conditions, however, show a different picture for the RP50 and RP65 criteria in particular. Under the 1,000 sample condition, the false positive rates are all below 5% for Performance Level 2, with the exception of RP 50 at 11%. For Performance Level 3, the false positive rate for RP65 was 11% and for RP 50 was 24%. For Performance Level 4, the false positive rate for RP80 was 6% and for RP 65 was 10%. For Performance Level 5 the false positive rates were all below 3%, except for RP80 at 6%.

Results across sample size conditions did not improve as sample size increased and continued to show a similar pattern. For example, under the 50,000 sample condition, for Performance Level 2, the false positive rate continued to be 11% for RP50. At Performance Level 3 the false positive rate for RP50 actually increased to 27%.

Similarly, the shape of the RP65 criterion results was consistent across sample sizes with relatively high false positive rates for Performance Levels 3 and 4.

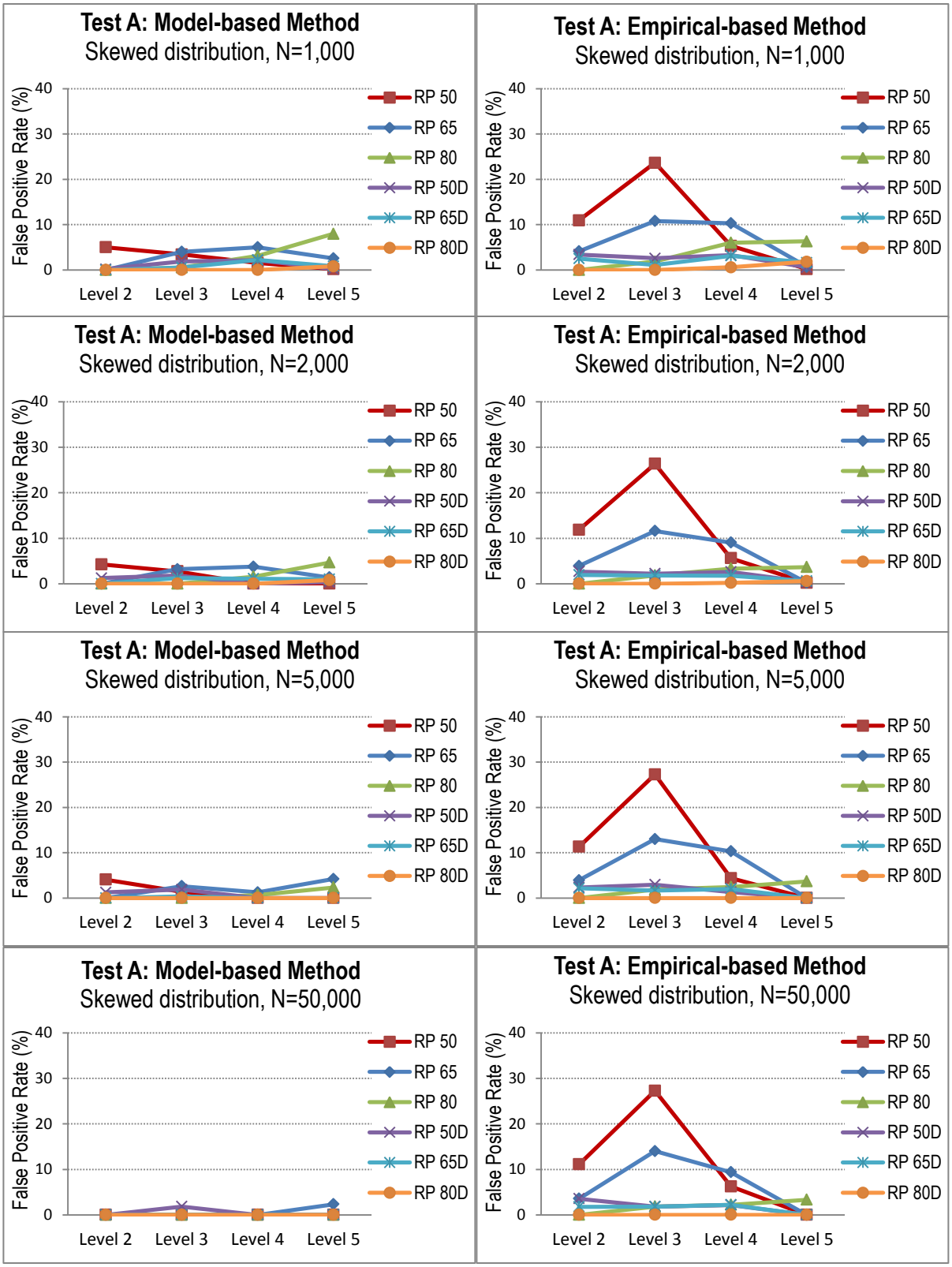

Figure 4.3.2 False Positive Results for Test A under Skewed Distribution Condition

Figures 4.3.3 and 4.3.4 display the True Positive Rate results for Test A. Figure 4.3.3 displays the True Positive Rate results for Test A under the normal distribution condition. Results are generally high for the Model-based method with one notable exception, namely the Performance Level 3 results for RP50D. Beginning with the Model-based method, under the 1,000 sample condition, results show that the true positive rate for RP50 for Performance Level 2 was 95%. There were no "true" items for the other item-mapping criteria and as such true positive results are not applicable. For Performance Level 3, the rates were all above 70% with the exception of RP50D whose rate was 20%. It is important to note that there are only two "true" exemplar items for RP50D at Performance Level 3. For Level 4, the true positive rates ranged between 69% (RP65D) and 95% (RP50). And finally for Performance Level 5, the true positive rates ranged between 78% (RP80D) and 98% (RP50).

Under the 2,000 sample condition and the 5,000 sample condition, the true positive rate results were very similar. Across all performance levels and item-mapping criteria, the true positive results ranged from 70% to 100% with the same notable exception for RP50D. For Performance Level 3, the true positive rate for RP50D was 5% and 15% for the 2,000 and 5,000 sample conditions respectively.

Under the 50,000 sample condition, the true positive rate results were generally higher, at or above 80% across all levels and item-mapping criteria with, once again, one exception: the true positive rate for Performance Level 3 and RP50D criterion was actually 0%. Further investigation on this and other unusual results will be presented in Chapter 5.

The results for the Empirical-based method were similar under the 1,000 sample condition to the Model-based results, but interestingly for Performance Level 4, results became worse as sample size increased. Under the 1,000 sample condition, results show that the true positive rate ranged between 5% (RP50D) and 85% (RP65 and RP80) for Performance Levels 3. For Performance Level 4, true positive rates ranged between 61% for RP65D and 93% for RP50. For Performance Level 5, the true positive rates were generally high, all at or above 87%.

True positive rate results under the 2,000 and 5,000 sample conditions were similar to one another. For Performance Level 2, the true positive rates were both 95% for RP50 under both sample size conditions. For Performance Level 3, the RP50D rate continued to be very low (15% under the 2,000 and 0% under the 5,000 condition). For Performance Level 4, the range in the true positive rates increased from the 1,000 condition. For the 2,000 condition the rates ranged between 43% and 92%. Similarly for the 5,000 condition, the rates ranged between 43% and 94%. For Performance Level 5, under the 2,000 condition, the true positive rates ranged from 72% for RP80D to 100% for RP50 and RP50D. Similarly for Performance Level 5, under the 5,000 sample condition, the true positive rates ranged from 75% for RP80D to 100% for RP50D.

Finally, under the 50,000 sample condition the true positive rates generally decreased for Performance Level 4. For Performance Level 2, the true positive rate was 100% for RP50. For Performance Level 3, the RP50D rate continued to be extremely low (actually 0%) but for the remaining item-mapping criteria the rates were at or above 83%. For Performance Level 4, the true positive rates ranged between 25% for RP80D and 96% for RP50. For Performance Level 5, the true positive rates were all at or above 83%.

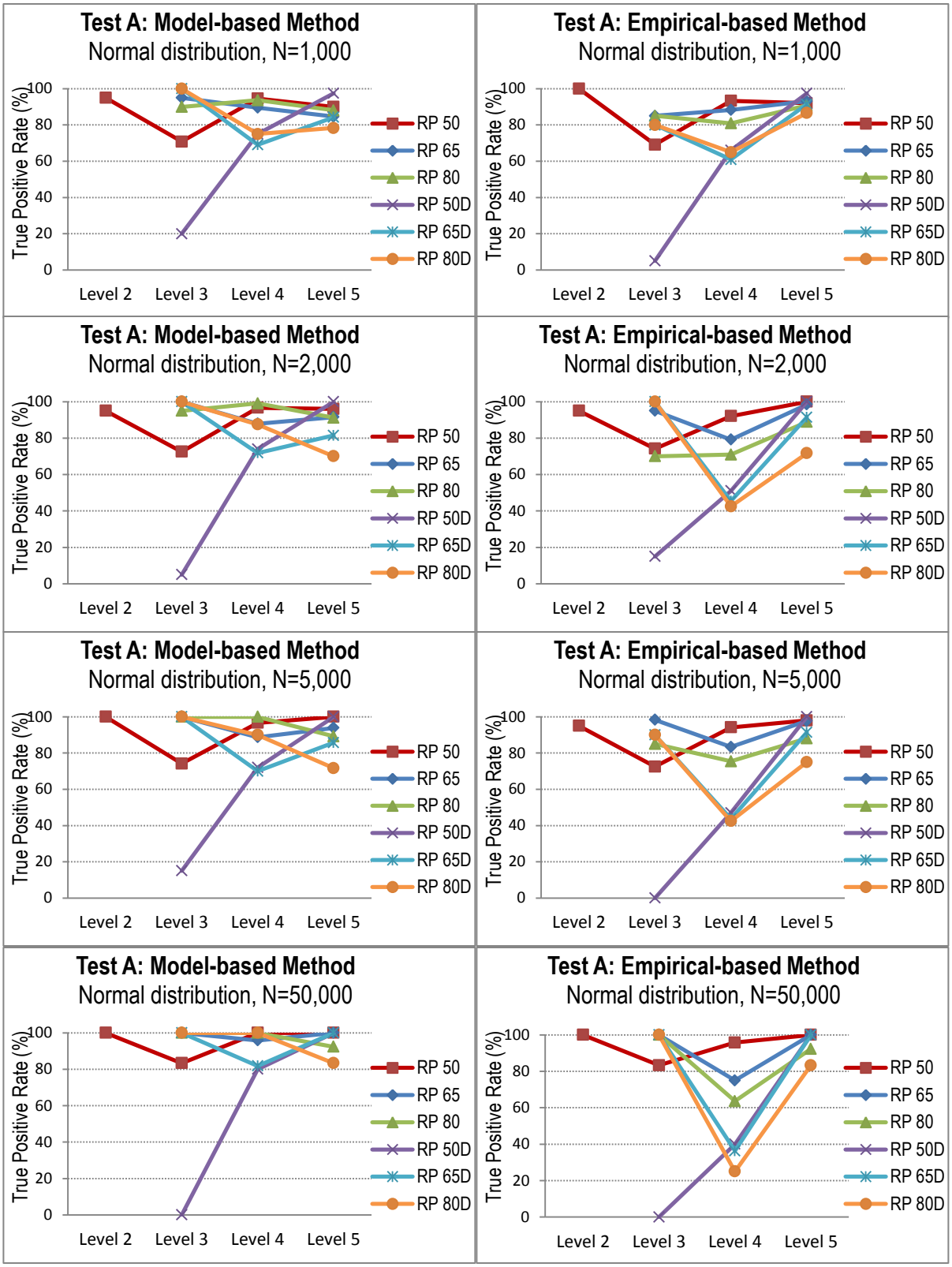

Figure 4.3.3 True Positive Results for Test A under Normal Distribution Condition

Figure 4.3.4 displays the True Positive Rate results for Test A under the skewed distribution condition. Beginning with the Model-based method, under the 1,000 sample condition, results show that the true positive rate for RP 50 for Performance Level 2 was 95%. For Performance Level 3, the rates were all above 77% again with the exception of RP 50D whose rate was 20%. For Level 4, the true positive rates ranged between 63% (RP65D) and 95% (RP80). And finally for Performance Level 5, the true positive rates ranged between 58% (RP80D) and 100% (RP50D).

Under the 2,000 sample condition and the 5,000 sample condition, the true positive rate results were once again very similar. Across all performance levels and item-mapping criteria, the true positive results ranged from 60% to 100% with the same notable exception for RP50D. For Performance Level 3, the true positive rate was 40% and 30% for the 2,000 and 5,000 sample conditions respectively.

Under the 5,000 sample condition, the true positive rate results were generally higher, at or above 80% across all levels and item-mapping criteria with the exception RP50D and RP65D at Level 4 with rates of 70% and 73% respectively. Noteworthy is the fact that the RP50D true positive rate for Level 3 was 100% under this condition.

The results for the Empirical-based method were different from the Model-based results, with true positive rates generally being lower and more variable across the board. Under the 1,000 sample condition, results show that the true positive rate ranged between 5% (RP50D) and 100% (RP80) for Performance Levels 3. For Performance Level 4, true positive rates also ranged widely from 25% for RP50D and RP65D to 89% for RP80. For Performance Level 5, the true positive rates ranged between 40% for RP80D and 79% for RP80.

True positive rate results under the 2,000, 5,000 and 50,000 sample conditions were similar to the 1,000 condition and notably did not improve as sample size increased. For example, under the 50,000 sample condition the true positive rate ranged between 0% (RP50D, RP65D, and RP80D) and 100% (RP80) for Performance Level 3. For Performance Level 4, true positive rates also ranged widely from 20% for RP50D to 91% for RP80. And finally for Performance Level 5, the true positive rates ranged between 33% for RP80D and 85% for RP80.

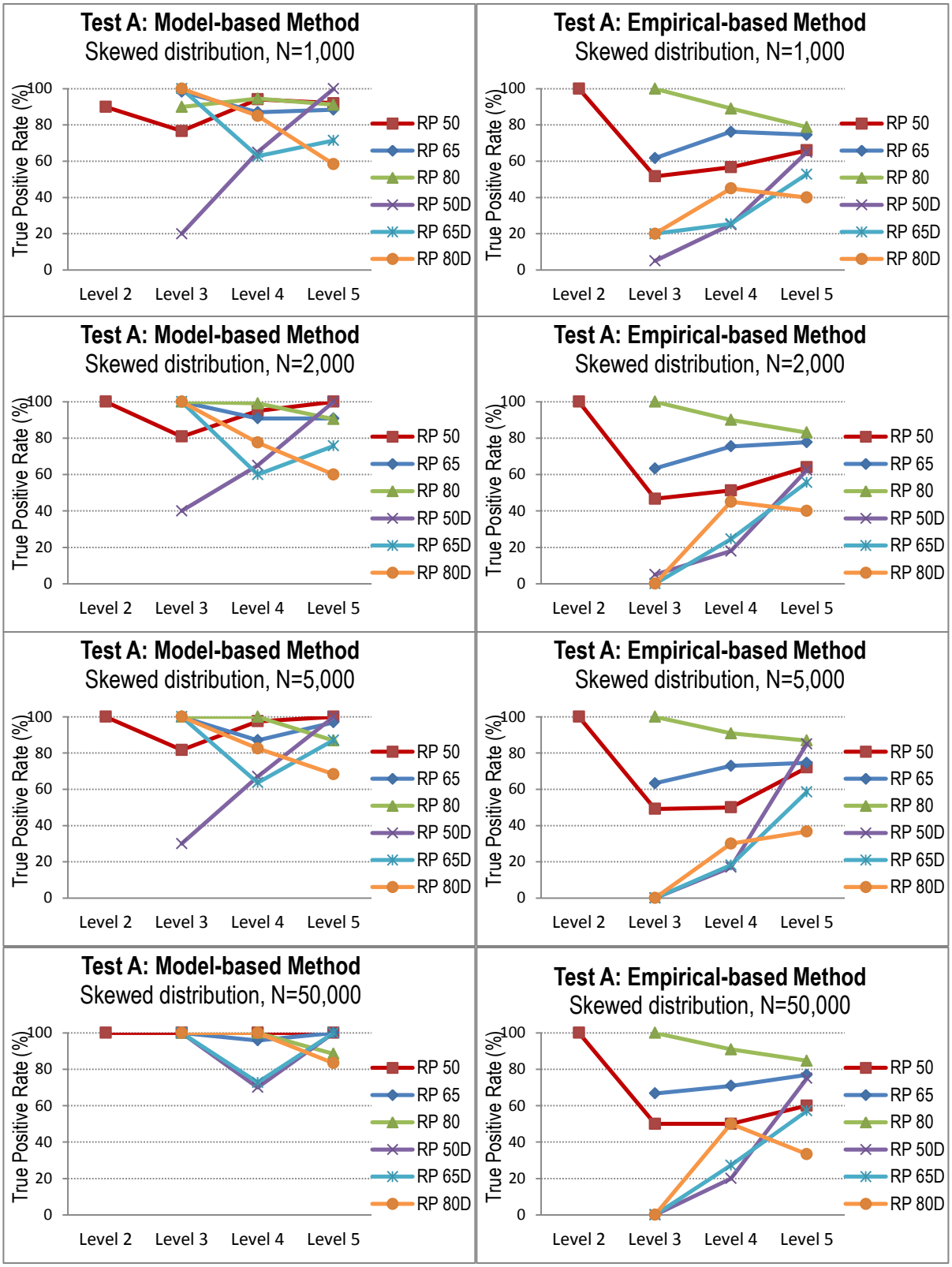

Figure 4.3.4 True Positive Results for Test A under Skewed Distribution Condition

4.4 Test B results

Figure 4.4.1 displays the False Positive Rate results for Test B under the normal distribution condition. Beginning with the Model-based method, the false positive results for Test A were slightly lower than for Test B under the normal distribution condition Under the 1,000 sample condition, the false positive rates for Performance Level 2 were all at or below 4%. For Performance Level 3, the false positive rates were all at or below 2%. For Performance Level 4, the false positive rates were all at or below 1%. And for Performance Level 5, the false positive rates were all 2% or lower.

Results for the 2,000, 5,000 and 50,000 sample conditions were similarly low. For example, for the 50,000 condition, the false positive rates for Performance Level 2 all 4% or lower. For Performance Levels 3, 4 and 5, the false positive rates were all at or below 2%. And for Performance Level 5, the false positive rates were all 2% or lower with the exception of RP80 with a 10% false positive rate.

The False Positive Rate results for the Empirical-based method under the normal distribution condition were very similar to those for the Model-based method with rates at or below 5% across all performance levels, item-mapping criteria, and sample sizes. Under the 1,000 sample condition, the false positive rates for Performance Level 2 were all at or below 4%. For Performance Level 3, 4, and 5 the false positive rates were all at or below 2%.

Under the 5,000 sample condition, the false positive rates for Performance Level 2 were all at or below 5%. For Performance Level 3, the false positive rates were all 0% with the exception of RP50 with a 4% rate. For Performance Level 4, the false positive

rates were all 0%. And for Performance Level 5, the false positive rates were all 2% or lower.

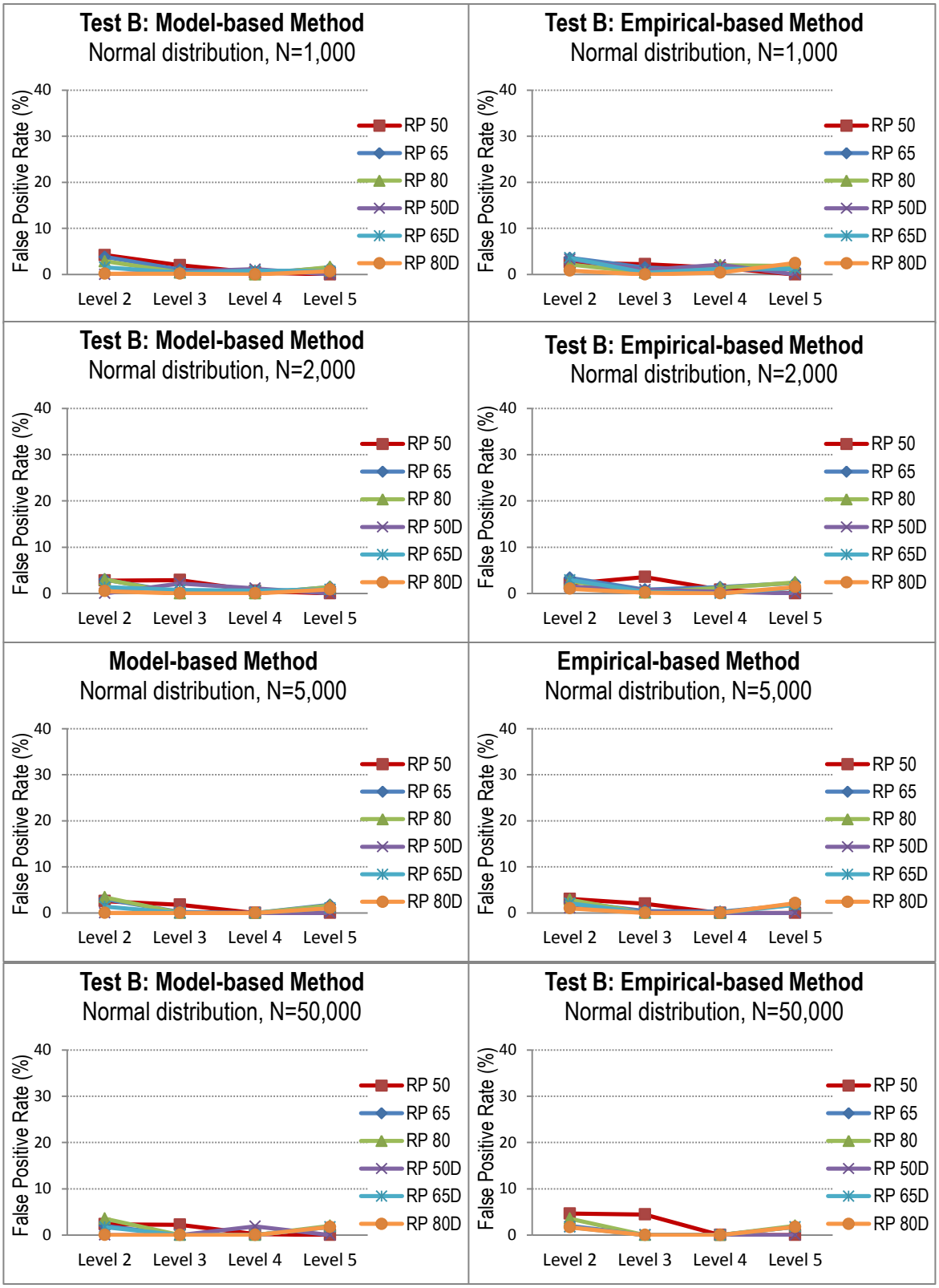

Figure 4.4.1 False Positive Results for Test B under Normal Distribution Condition

Figure 4.4.2 displays the False Positive Rate results for Test B under the skewed distribution condition. For the Model-based method, results are almost identical to those under the normal distribution condition. Under the 1,000 sample condition, the false positive rates for Performance Levels 2 and 3 were all at or below 4%. For Performance Level 4, the false positive rates were all at or below 1%. And for Performance Level 5, the false positive rates were all 2% or lower.

The Model-based results under the skewed distribution improved slightly with the increase in sample size. As such, the results for the 50,000 condition were as follows. The false positive rates for Performance Level 2 were all 4% or lower. For Performance Levels 3 and 4 the false positive rates were all 0%. And for Performance Level 5, the false positive rates were all 2% or lower.

For the Empirical-based Method the results were different under the skewed distribution condition compared to the Model-based method. Specifically results were worse for Performance Levels 2 and 3 but better for Performance Level 5. Under the 1,000 sample condition, the false positive rates for Performance Level 2 ranged from 3% for RP80D to 14% for RP50. For Performance Level 3, the false positive rates ranged from 3% for RP80D to 12% for RP65. However, for Performance Level 4 the rates were lower, all at or below 3%. And for Performance Level 5, all the false positive rates were 0%.

For the Empirical-based method under the skewed distribution condition, results did not improve as sample size increased. As such, the results for the 50,000 sample size condition, for example, were almost identical. Under the 50,000 sample condition, the false positive rates for Performance Level 2 ranged from 2% for RP80D to 12% for

RP50. For Performance Level 3, the false positive rates ranged from 2% for RP80D to 13% for RP50. However, for Performance Level 4 the rates were lower, all at or below 2%. And for Performance Level 5, all the false positive rates were 0%.

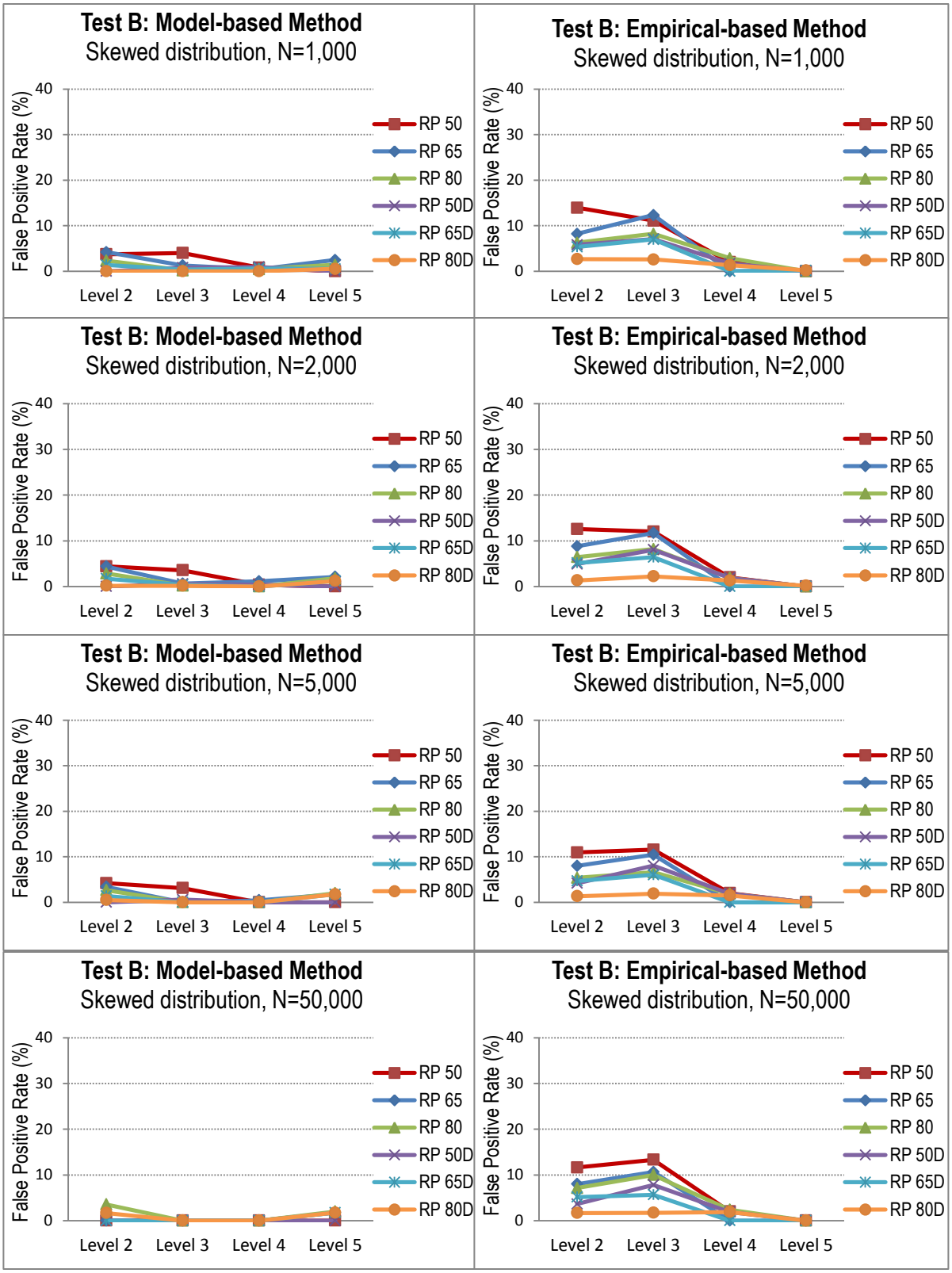

Figure 4.4.2 False Positive Results for Test B under Skewed Distribution Condition

Figures 4.4.3 and 4.3.5 display the True Positive Rate results for Test B. Figure 4.4.3 displays the True Positive Rate results for Test B under the normal distribution condition. Beginning with the Model-based method, results show a wide range in the true positive rates with low rates found mostly for the item-mapping criteria that included the discrimination factor. Under the 1,000 sample condition, the true positive rates for Performance Leve 2 were 95% or greater for RP50, RP65 and RP80, but only 46% for RP50D and 10% for RP65D. For Level 3, the true positive rates were 84% or greater for RP50, RP65 and RP80, but 62% for RP50D, 53% for RP65D and 35% for RP80D. For Level 4, true positive rates were 71% or above for all item-mapping criteria. For Level 5, the true positive rates were generally higher for most criteria (at or above 92%) with the exception of RP80D with a rate of 73%.

Of note is that these results did not improve as sample size increased. As such, even with the 50,000 sample, the item-mapping criteria with a discrimination factor generally performed worse than the criteria without the discrimination factor. As noted previously, further investigation into these results will be discussed in Chapter 5.

The results for the Empirical-based method under the normal distribution condition were also unexpected. Under the 1,000 sample condition, for Performance Level 2, all true positive values were at or above 70%. For Performance Level 3, the true positive rate ranged between 20% and 89% with the item-mapping criteria with discrimination performing worse than those without. For Performance Level 4, the true positive rates were all at or above 77%. And for Performance Level 5 all true positive rates were high, at or above 93%.

True positive rate results under the 2,000, and 5,000 sample conditions were similar to one another with the lowest rates occurring at Performance Levels 3 and 5 for the RP80D criterion. Under the 50,000 condition, results improved slightly. For Level 2, all true positive rates were at or above 80%. At Level 3, RP50, RP65, and RP80 performed higher (at or above 80%) than RP50D, RP65D, and RP80D (56%, 57%, and 50% respectively). At Performance Level 4, the rates ranged between 57% for RP50D and 94% for RP65 and RP80. And finally for Performance Level 5, the true positive rates were all 100% except for RP80 at 90% and RP80D at 67%.

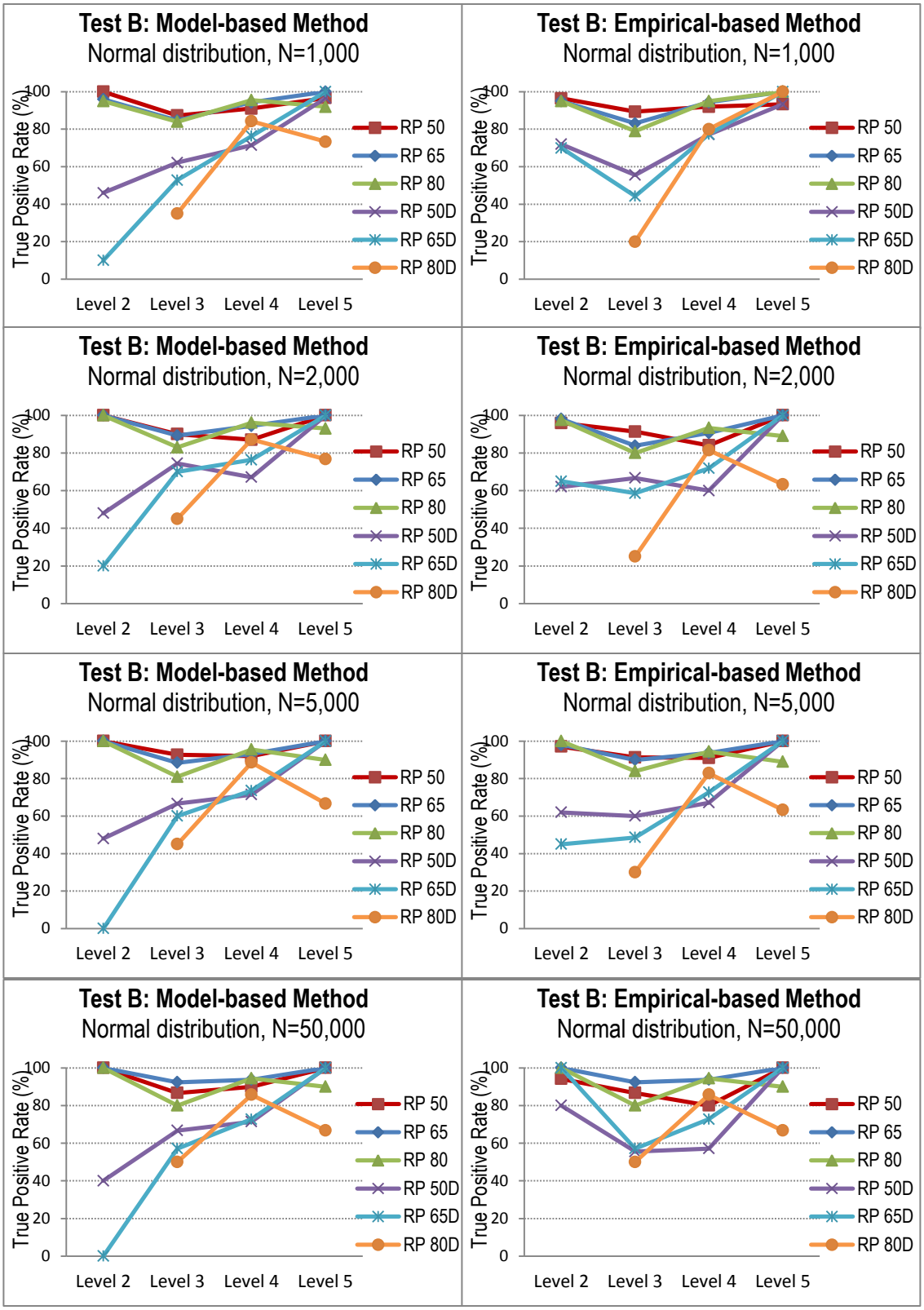

Figure 4.4.3 True Positive Results for Test B under Normal Distribution Condition

Figure 4.4.4 displays the True Positive Rate results for Test B under the skewed distribution condition. For the Model-based method, the results for the skewed distribution condition were very similar to the normal distribution condition. Beginning with the Model-based method, results show a wide range in the true positive rates with lower rates found mostly for the item-mapping criteria that included the discrimination factor. Under the 1,000 sample condition, the true positive rates for Performance Level 2 were 97% or greater for RP50, RP65 and RP80, but only 42% for RP50D and 10% for RP65D. For Performance Level 3, the true positive rates were 85% or greater for RP50, RP65 and RP80, but 56% for RP50D, 49% for RP65D and 50% for RP80D. For Level 4, true positive rates were 82% or above for RP50, RP65 and RP80, but 57% for RP50D, 69% for RP65D and 83% for RP80D. For Level 5, the true positive rates were 100% for RP50, RP65, RP50D and RP65D, 82% for RP 80 and only 40% for RP80D.

Of note is that these results did improve in many cases as sample size increased. As such, the true positive results under the 50,000 sample were as follows. True positive rates for Performance Level 2 were 100% for RP50, RP65 and RP80, 80% for RP50D, and 50% for RP80D. For Level 3, the true positive rates were generally high, all at or above 78%. Similarly for Level 4, the true positive rates were generally all at or above 71%. For Level 5, the true positive rates were 100% for RP50, RP65, RP50D and RP65D, 80% for RP 80 and only 33% for RP80D.

The results for the Empirical-based method under the skewed distribution condition were also unexpected and not necessarily better than the Model-based results. Under the 1,000 sample condition, the true positive rates for Performance Level 2 were 100% for RP65 and RP80, 94% for RP50, but only 52% for RP50D and 40% for RP65D.

For Level 3, the true positive rates were 65% for RP50 and RP80, 68% for RP65, 50% for RP50D, 37% for RP65D and only 15% for RP80D. For Level 4, true positive rates were 50% for RP50, 63% for RP65, 77% for RP80, 47% for RP50D, 35% for RP65D and 26% for RP80D. For Level 5, the true positive rates were higher for most criteria (at or above 86%) with the exception of RP50 and RP50D with rates of 67%.

Similar to the Model-based method, the true positive rate results under the skewed distribution condition did not appear to improve as sample size increased and in fact worsened in a few cases. As such, the results under the 50,000 were similarly variable and with generally lower rates for the item-mapping criteria with a discrimination factor. Under the 50,000 sample condition, the true positive rates for Performance Level 2 were 100% for RP65 and RP80, 88% for RP50, but only 40% for RP50D and 0% for RP65D. For Level 3, the true positive rates were 67% for RP50, 69% for RP65, 60% for RP80, 44% for RP50D, 29% for RP65D and 0% for RP80D. For Level 4, true positive rates were 40% for RP50, 69% for RP65, 72% for RP80, 43% for RP50D, 27% for RP65D and only 14% for RP80D. For Level 5, the true positive rates were higher for most criteria (at or above 90%) with the exception of RP50 and RP50D with rates of 67%.

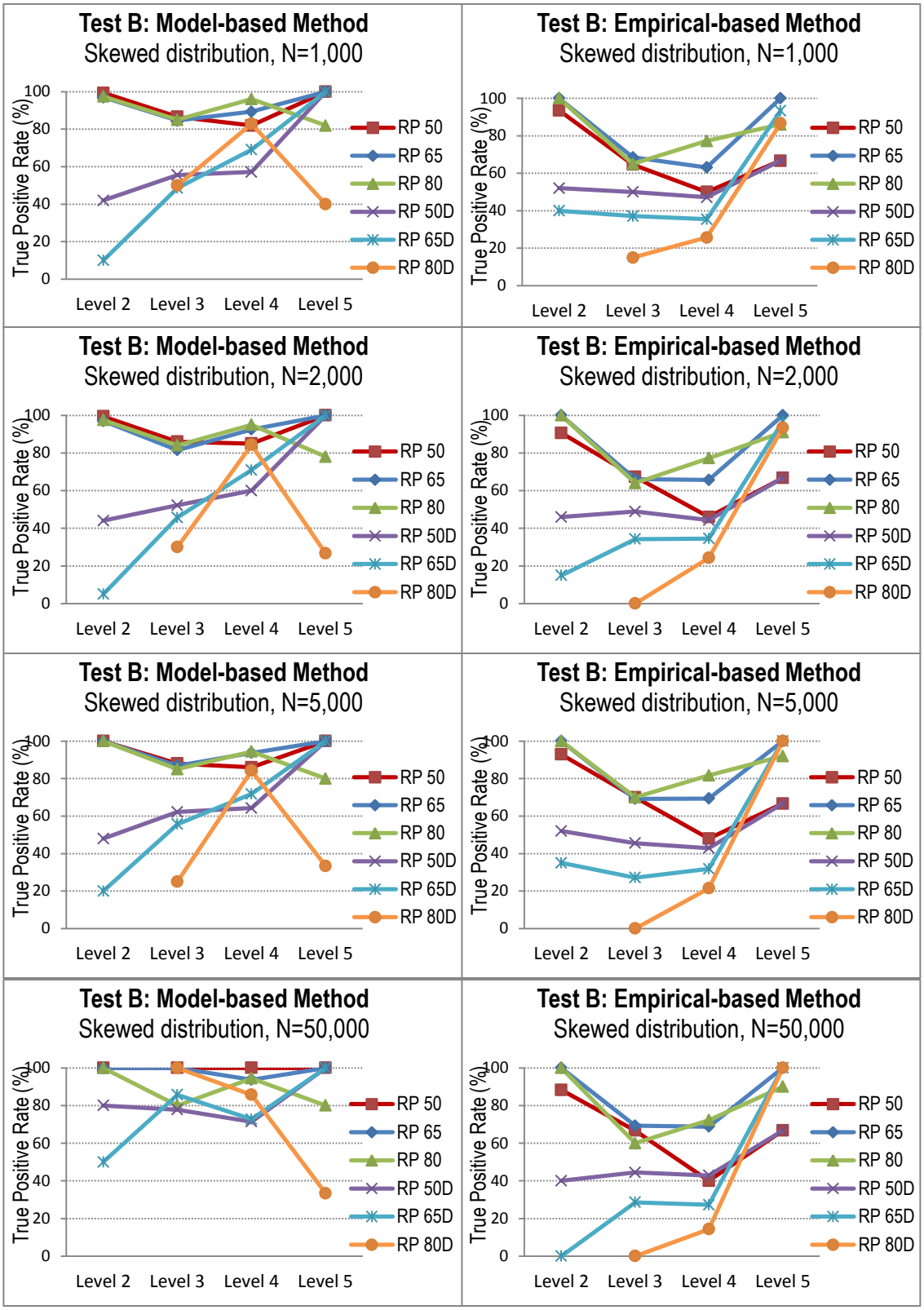

Figure 4.4.4 True Positive Results for Test B under Skewed Distribution Condition

4.5 Criteria for Selecting Exemplars: Summary of Results

Because the results for both Test A and B often differed across Performance Levels and Item-mapping Methods an attempt was made below to synthetize the results for each criterion for selecting exemplars. This was done by aggregating results across the four sample sizes, the two item-mapping methods, and across the four performance levels. The reason for doing so is that ultimately one methodology will have to be chosen if/when a State decides to use item-mapping for identifying exemplars.

Table 4.5.1 presents a summary of the False Positive Rate results for each test, ability distribution and criterion for selecting exemplars. For example, under the normal condition for Test A, 81% of the false positive rates across all sample sizes and performance levels were less than or equal to .05. The table shows that for both tests the results under the normal distribution condition are generally better than the results under the skewed distribution condition, with one exception for Test A using RP80. The table also shows that using a discrimination criterion in addition to using RP alone, improved the false positive rate results. In most instances, 100% of the false positive rates for the item mapping methods using the discrimination criteria were at or below the .05 threshold. Of note, RP80D performed the best across both tests and ability distributions and RP50 performed the worst across both tests and ability distributions.

Table 4.5.1 Summary of False Positive Rate Results

% False Positive Rate ≤ .05	Test A		Test B	
	Normal	Skewed	Normal	Skewed
RP 50	81%	63%	100%	75%
RP 65	84%	72%	100%	75%
RP 80	81%	91%	100%	75%
RP 50 + Disc.	100%	100%	100%	84%
RP 65 + Disc.	100%	100%	100%	78%
RP 80 + Disc.	100%	100%	100%	100%

Table 4.5.2 similarly summarizes the True Positive Rate results for each test, ability distribution, and criterion for selecting exemplars. For example, under the normal condition for Test A, 81% of the true positive rate results across all sample sizes and performance levels were greater than or equal to .80. Once again, the table shows that for both tests the results under the normal distribution condition were better than the results under the skewed distribution condition with the exception of RP80 for Test A. On the other hand, the table shows that using a discrimination criterion in addition to using RP alone, decreased the true positive rate results considerably. Unlike the false positive rate results, no single criterion for selecting exemplars performed better than the others across both tests and ability distributions although RP50D consistently performed the worst. More importantly, the true positive rates in many cases were very low, particularly for Test B.

Table 4.5.2 Summary of True Positive Rate Results

% True Positive Rate ≥ .80	Test A		Test B	
	Normal	Skewed	Normal	Skewed
RP 50	81%	59%	100%	63%
RP 65	97%	63%	100%	75%
RP 80	88%	97%	97%	75%
RP 50 + Disc.	53%	44%	28%	16%
RP 65 + Disc.	78%	44%	28%	28%
RP 80 + Disc.	69%	50%	53%	53%

The results presented in sections 4.3 and 4.4 were presented at a "zoomed in" level with results for each test, performance level, item-mapping method, criterion for selecting exemplars, ability distribution and sample size presented. At this "zoomed in" level of analyzes, no clear findings were found as to which item-mapping method or criteria for selecting exemplars was systematically better across all simulated conditions and across the four performance levels. Section 4.5 attempted to "zoom out" a little to see if there were any patterns in the findings by aggregating results across the different performance levels, item-mapping methods and sample sizes. The next chapter will provide a more detailed summary of results, provide hypotheses regarding some of the unexpected results, and attempt to derive any implications or recommendations from these findings.

CHAPTER 5

DISCUSSION

5.1 Overview

This study was designed to examine the performance of several item-mapping methods for identifying exemplars. A Monte Carlo simulation study was conducted to examine the performance of both empirical and model-based methods, of different criteria for selecting items (RP alone or RP and discrimination combined), of different RP values (50, 65, 80), of different sample sizes (1,000, 2,000, 5,000, 50,000), and of different ability distributions (normal or skewed). The simulation study was designed to mimic, to the extent possible, two statewide assessments. Specifically, operational data from a 2002 administration of a Grade 10 science assessment and a 2006 administration of a Grade 10 mathematics assessment were used to generate the simulated data in the study, and were also used as a guide for choosing the different simulation conditions. Results of the various item mapping methods and simulation conditions were evaluated in terms of both False Positive Rate (how often were the incorrect exemplars identified) and True Positive Rate (how often were the correct exemplars identified).

This chapter discusses the major findings of the study with reference to the existing literature where applicable. The first section discusses the results for the different item mapping methods and simulation conditions. Next, further investigations into the possible reasons for the unexpected results are presented. This is followed by a discussion on the limitations of the study and directions for future use. The chapter concludes with outlining some implications and recommendations regarding the use of item mapping methods for the purpose of selecting exemplars.

5.2 Summary of Findings

As described in more detail below, this study found no definite findings regarding which item-mapping method or criteria for selecting exemplars was systematically better across all simulated conditions and across the four performance levels. A few interesting findings were observed, however, and will be highlight below.

5.2.1 Method Type

For Test A, the Model-based method performed better than the Empirical-based Method both in terms of true positive and false positive rates. This finding held true for both the normal and skewed distribution conditions. This finding was not found for Test B. The false positive rates for the Empirical-based method and Model-based method were similar under the normal distribution condition. The Model-based method did perform better than the Empirical-based Method under the skewed distribution condition in terms of false positive rates. With regards to the true positive rates for Test B neither model performed particularly well specially for the item mapping criteria that included a discrimination factor.

5.2.2 Shape of Ability Distribution

With respect to both true positive and false positive rates, results under the normal distribution condition appeared better than under the skewed distribution condition for the Empirical-based method but no clear patterns were observed between the two distributions for the Model-based method, suggesting that after rescaling, the Model-based method may be less susceptible to changes in the shape of the distribution than the Empirical-based method.

5.2.3 Criteria for Selecting Exemplars

There was no clear pattern across all conditions and performance levels as to which criteria used to select exemplars performed better; however a few patterns were observed. Using a discrimination criteria in addition to using RP alone, improved the false positive rate results for both tests. The converse was true, however, for the true positive rate results. Results showed that using a discrimination criterion in addition to using RP alone, decreased the true positive rate results. In terms of false positive rates, RP80D performed the best across both tests and ability distributions and RP50 performed the worst across both tests and ability distributions. In terms of true positive results, RP50D consistently performed the worst.

5.2.4 Sample Sizes

With regards to sample size, the true and false positive results generally decreased for both tests as sample size increased for the Model-based method but remained generally unchanged for the Empirical-based method. Where results did improve for the Model-based method, the improvements were generally small, however.

5.2.5 Further Investigation

As allude to earlier, the true positive results for many of the conditions in this study were unexpectedly low particularly for Test B which warranted further investigation. First, select items from each test, those that appeared to consistently result in low true positive rates for particular Performance Levels, were chosen for further review. An example from each test is presented below in turn. And second, the relationship between the item parameters estimates and the "true" item parameters were examined to rule out any potential issues with the calibration and scaling of the items.

Results for the 50,000 sample replications are presented in Appendix C for illustrative purposes.

Results presented in Chapter 4 showed that, for Test A, RP50D performed consistently poorly at Performance Level 3; therefore, these results were chosen for further investigation. For Test A, the RP50D method resulted in two "true" exemplars at Performance Level 3. The two exemplar items were items #21 and item #40, both multiple-choice items and which will be further examined in turn.

The true IRT parameters for item #21 were as follows: a=1.333; b=-.888; and c=.044. The Response Probability (RP) at the midpoint of Performance Level 3 (the performance level of interest) was .89. The discrimination associated with Performance Level 3 was .43 (difference between RP at midpoint of level 3 (RP=.89) and midpoint of level 2 (RP=.46)). As such, item #21 was identified as an exemplar because the RP value is greater than .50 and the discrimination is greater than .30. Note that this item was not identified as an exemplar for Performance Level 2 because the RP value associated with Level 2 (RP=.46) was just short of the .50 criterion. Now let us examined what happened under the Model-based method condition under the normal distribution and 50,000 sample condition (a condition where one would expect sufficient sample size to estimate the parameters in a stable manner). The IRT parameters for item #21 after scaling were as follows: a= 1.287; b=-.951; and c=.061. The Response Probability at the midpoint of Performance Level 3 was .90. The discrimination associated with Performance Level 3 was .39 (difference between RP at midpoint of level 3 (RP=.90) and midpoint of level 2 (RP=.51)). As such, item #21 would have been identified as an exemplar because the RP value is greater than .50 and the discrimination is greater than .30, except for the rule that

an item can only be an exemplar for one performance level. In this replication, Item #21 was identified as an exemplar for Performance Level 2 since it met the .50 RP and the .30 discrimination threshold and thus was no longer eligible to be identified for Performance Level 3. In other words, a very small difference in the item parameter estimates resulted in item #21 being identified as an exemplar for Performance Level 2 but not Performance Level 3.

The other "true" exemplar for Performance Level 3 using the RP50D criterion was item #40. The true IRT parameters for item #40 were as follows: a= .928; b=-.179; and c=.134. The Response Probability (RP) at the midpoint of performance level 3 was .63. The discrimination associated with Performance Level 3 was .31. As such, item #40 was identified as an exemplar because the RP value is greater than .50 and the discrimination is greater than .30. Now let us examined what happened under the Model-based method condition under the normal distribution and 50,000 sample conditions. The IRT parameters for item #40 after scaling were as follows: a= .934; b=-.156; and c=.163. The Response Probability at the midpoint of Performance Level 3 was .63. The discrimination associated with Performance Level 3 was .296. As such, item #40 was not identified as an exemplar because the discrimination was just below .30. Note, that in this instance different rounding rules would have resulted in different outcomes. Again, a very small difference in the item parameter estimates resulted in item #40 not being identified as an exemplar for performance level 3 based on the RP50D criterion.

Results presented in Chapter 4 showed that, for Test B, RP65D performed consistently poorly at Performance Level 2; therefore, these results were also chosen for further investigation. For Test B, the RP65D method resulted in two "true" exemplars at

Performance Level 2. The two exemplar items were items #10 and item #20, both multiple-choice items and which will be further examined in turn.

The true IRT parameters for item #10 were as follows: a=.779; b=-1.730; and c=0. The Response Probability (RP) at the midpoint of Performance Level 2 was .72. The discrimination associated with Performance Level 2 was .31. As such, item #21 was identified as an exemplar because the RP value is greater than .65 and the discrimination is greater than .30. Once again we will compare theses "true" values to the values obtained from the Model-based method, normal distribution, 50,000 sample condition. The IRT parameters for item #10 after scaling were as follows: a= .737; b=-1.782; and c=.053. The Response Probability at the midpoint of Performance Level 2 was .74. The discrimination associated with Performance Level 2 was .28. As such, item #10 was not identified as an exemplar the discrimination was lower than .30.

The other Test B "true" exemplar for Performance Level 2 using the RP65D criterion was item #20. The true IRT parameters for item #20 were as follows: a= 1.042; b= -1.61; and c=.183. The RP at the midpoint of performance level 2 was .79. The discrimination associated with Performance Level 2 was .34. As such, item #20 was identified as an exemplar because the RP value is greater than .65 and the discrimination is greater than .30. Under the Model-based method condition under the normal distribution and 50,000 sample conditions, the IRT parameters for item #20 after scaling were as follows: a= .994; b= -1.651; and c=.250. The Response Probability at the midpoint of Performance Level 2 was .81. The discrimination associated with Performance Level 2, however, was .295. As such, item #20 was not identified as an exemplar because the discrimination was just below .30. Note that once again different

rounding rules would have resulted in different outcomes. Again, a small difference in the item parameter estimates resulted in item #20 not being identified as an exemplar for performance level 2 based on the RP65D criterion.

The close look at select items suggested that small differences in the item parameters could result in low true positive rates for items where the RP or discrimination were borderline but did not explain why true positive results were so low in some cases. To rule out any issues with the calibration or scaling of the item parameters, aa-plots, bb-plots, and cc-plots were generated for a number of replications. Appendix C displays theses plots for the 50,000 conditions. The aa-plots and bb-plots showed no concerns with the calibration or the scaling but the cc-plots revealed a potential problem with both the generation and calibration of the c parameters. The "true" item parameters were obtained from operational technical manuals and some of the c values were reported as zero. When the data was generated for the study, those zero values were assumed to be true. My hypothesis is that those c values may have been fixed for the operational assessments because the items did not converge properly. It is common practice to fix the c parameters in those instances although some programs fic the c parameters to zero and others fix the c parameters to the probability of random guessing (.25 for 4 option multiple-choice questions). When the data was generated for this study and then the items calibrated, it is apparent that the c values for these items were not estimated correctly. This is may be due to how WinGen and Parscale generate initial values for the c parameter estimates. Regardless, the cc-plots show that the c values do not correlate as highly as expected between the true values and values obtained from the different calibrations, even when the sample size is as large as 50,000. This may

be the reason why small differences in item parameter estimates resulting in low true positive rates can be found in this study and appear to be slightly biased. When the c-estimate is higher than it should be, it tends to depress the a-estimate. The fact that Test B had more of these problematic c parameters, may be the reason why results showed lower true positive rates for Test B, particularly for the item mapping methods that relied on the discrimination value.

5.3 Limitations and Directions for Future Research

As discussed in the previous section, the estimation of the c parameters for a number of multiple-choice items (those with a c value of zero) may have depressed true positive rates in this study. Future research should attempt to generate the data as a 2PL model to see whether the true positive results do indeed improve for both tests, but in particular for Test B.

For this study item response data were generated under a no model misfit condition, IRT item parameter estimates from the operational statewide assessment mentioned above were used as generating item parameter values. Item response data for the multiple-choice items were generated using the three-parameter logistic model (3PLM). Data for the short-answer items were generated using the two-parameter logistic model (2PLM). Lastly, data for the extended-response items were generated using the generalized partial credit model (GPCM). As such, the data were generated from the models and can therefore be assumed to fit the models fully. This condition provided a "best case" scenario that could be used for comparing model-based item-mapping methods; however, it remains an unrealistic condition. Future research should make an attempt to generate realistic model misfit perhaps by estimating non-parametric item

response functions. This could be done by using a Kernel smoothing method (Ramsay, 1991).

Additional limitations have to do with the limited scope and thus generalizability of this study. Although the study examined several different variables, ultimately only two statewide tests served to guide the simulation conditions. Furthermore, operationalizing any of the item mapping methods required numerous decision points all of which may have resulted in different findings. For example, in this study the discrimination criterion was based on a 30 point difference in response probability between adjacent performance levels. Would the results be different if a 25 or 35 point difference was chosen? Another example, in this study a rule was applied that an item could only map to one Performance Level. Although this rule was based on the literature review, results from this study suggest that this rule may be too strict and lead to decreased true positive rates. This and many other decision points put in question the generalizability of the study results beyond the two tests examined.

This study relied solely on simulated data and "true" exemplars were determined by the item response data without regards for the actual content of the items. Ultimately, whether an item represents a true exemplar should be determined by a content review of the items themselves, the standards they purport to measure, and the performance level descriptors associated with each assessment. Future research should make an attempt to examine whether different item-mapping methods are better than others at identifying exemplars that have been previously identified by content experts.

5.4 Conclusion

Score reporting continues to be one of the most important and often neglected aspects of educational testing. The good news is that "efforts to communicate test results in clear and meaningful ways has recently become a higher priority for many testing agencies" (Zenisky & Hambleton, 2012, p.21). The bad news is that "score reporting is among the most challenging aspects of test development facing [those] testing agencies today" (Zenisky & Hambleton, 2013, p.175). Of particular importance is to provide users (students, parents, teachers, policy makers) with better descriptions of what each of the performance levels typically reported on statewide assessments actually mean. The use of item-mapping to select exemplar items has been taunted as a promising mechanism to provide that meaning for performance level scores; yet the research on this area remains insufficient.

This study attempted to add to the body of literature by examining the performance of two item-mapping methods and different criteria for identifying exemplars under several simulated conditions. The results of the study were neither clear nor systematic across all conditions and performance levels. What then are the implications and recommendations, if any, regarding the use of item mapping methods for the purpose of selecting exemplars based on this study? If a state wants to use an item-mapping method to select exemplars, which method/criteria should the state employ? As with most psychometric questions, this study suggests the answer is "it depends." This study suggests that a number of factors should be examined before choosing a methodology.

First, how many items does the state want to release and how many are available for release? By definition, using RP alone will result in more exemplars identified than using RP and discrimination combined.

Second, how difficult is the assessment? And how are the items distributed across the scale? As we saw in this study, there were instances where certain RP criteria did not result in any exemplars selected.

Third, how much time and resources are available for content experts to review the items selected via the item-mapping methods? If resources are vast, it may be reasonable to provide panelists with results from both item-mapping methods, as in the SAT study discussed earlier (Hambleton, Sireci & Huff, 2008), and allow panelists the freedom to make their own judgements. If resources are scarce, however, more strict apriori methodology and criteria may need to be applied so that panelists have time to review all items.

Fourth, how is the ability distribution distributed? If the theta distribution is skewed, this study suggests that fewer "true" exemplars may be identified using the Empirical-based model and as such those results should be used with caution.

This leads us to the last and final recommendation. Results from any item-mapping methodology remains a purely statistical, content-free, estimate of the type of item students at different performance levels are likely to be able to answer correctly. It may be a useful tool to narrow down the number of items that content specialists will need to review in search of exemplars but in no way replaces or diminishes the hard work of actually reviewing those items in relation to the standards and performance levels they

purport to measure and as such state agencies should continue to allocate the resources needed to conduct such work.

APPENDIX A

TRUE ITEM PARAMETERS AND EXEMPLARS

Table A.1 Test A IRT Item Parameters

Item #	Type	A	B	C		
1	MC	0.813	-0.773	0.094		
2	MC	1.162	0.22	0.475		
3	MC	0.834	1.055	0.071		
4	MC	0.936	0.728	0.119		
5	MC	0.851	0.582	0.198		
6	MC	1.46	0.683	0.205		
7	MC	0.991	0.813	0.205		
8	MC	0.892	0.041	0.35		
9	MC	1.071	1.085	0.287		
10	MC	1.309	0.893	0.275		
11	MC	0.978	0.607	0.222		
12	MC	1.647	0.455	0.194		
13	MC	0.796	0.09	0.255		
14	MC	1.356	0.87	0.122		
15	MC	0.575	0.33	0.094		
16	MC	1.049	1.161	0.192		
17	MC	0.73	-0.273	0.16		
18	MC	0.437	-0.69	0		
19	MC	1.496	0.452	0.172		
20	MC	0.348	-0.975	0		
21	MC	1.333	-0.888	0.044		
22	MC	1.175	0.699	0.257		
23	MC	0.72	0.062	0.152		
24	MC	0.682	-1.261	0		
25	MC	1.065	0.327	0.224		
26	MC	0.948	0.96	0.268		
27	MC	0.617	0.856	0.155		
28	MC	0.478	0.575	0.13		
29	MC	0.823	0.26	0.042		
30	MC	0.814	0.35	0.166		
31	MC	1.311	0.137	0.147		
32	MC	0.636	0.791	0.144		
33	MC	0.835	0.562	0.202		
34	MC	1.333	1.108	0.216		
35	MC	0.782	-0.038	0.065		
36	MC	2.122	1.184	0.123		
37	MC	0.444	0.434	0.113		
38	MC	0.854	0.451	0.135		
39	MC	0.775	1.328	0.141		
40	MC	0.928	-0.179	0.134		
Item #	**Type**	**A**	**D1**	**D2**	**D3**	**D4**
41	ER	1.287	-0.873	0.047	0.876	1.334
42	ER	1.229	-0.778	0.034	0.666	1.125
33	ER	1.11	-0.738	-0.043	0.592	1.325
44	ER	0.975	-0.768	0.029	1.303	2.167

Table A.2 “True” Exemplars for Test A

Item	RP 50				RP 65				RP 80				RP 50 D				RP 65 D				RP 80 D			
#	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5
1		✓				✓					✓													
2	✓					✓					✓													
3			✓					✓				✓												
4			✓				✓					✓			✓				✓					
5			✓				✓					✓												
6			✓				✓					✓			✓				✓					
7			✓				✓					✓												
8		✓				✓					✓													
9			✓					✓				✓												
10			✓				✓					✓			✓				✓					
11			✓				✓					✓			✓				✓					
12			✓				✓				✓				✓				✓				✓	
13		✓					✓				✓													
14			✓					✓				✓			✓					✓				✓
15			✓				✓					✓												
16			✓					✓				✓								✓				✓
17		✓				✓					✓													
18		✓					✓					✓												
19			✓				✓				✓				✓				✓				✓	
20		✓					✓					✓												
21		✓				✓				✓				✓				✓				✓		
22			✓				✓					✓			✓				✓					
23		✓					✓					✓												
24	✓					✓				✓														
25		✓					✓				✓								✓				✓	
26			✓				✓					✓												
27			✓					✓				✓												
28			✓					✓																
29			✓				✓					✓			✓				✓					
30			✓				✓					✓												
31		✓					✓				✓								✓				✓	
32			✓					✓				✓												
33			✓				✓					✓												
34			✓					✓				✓								✓				✓
35		✓					✓				✓													
36				✓				✓				✓				✓				✓				✓
37			✓				✓																	
38			✓				✓					✓			✓				✓					
39				✓				✓																
40		✓					✓				✓			✓										
41_1																								
41_2																								
41_3																								
41_4				✓				✓				✓				✓				✓				✓
42_1																								
42_2																								
42_3																								
42_4				✓				✓				✓				✓				✓				✓
43_1																								
43_2																								
43_3																								
43_4				✓				✓								✓				✓				
44_1																								
44_2																								
44_3																								
44_4																								

Table A.3 Test B IRT Item Parameters

Item #	Type	A	B	C		
1	MC	0.57	-2.63	0		
2	MC	0.632	-2.136	0.086		
3	MC	1.433	0.124	0.201		
4	MC	0.501	-1.076	0.176		
5	MC	0.754	-0.266	0.105		
6	MC	1.69	0.343	0.279		
7	MC	1.369	0.809	0.356		
8	MC	1.674	0.225	0.242		
9	MC	0.556	-1.63	0		
10	MC	0.779	-1.73	0		
11	MC	1.002	-0.204	0.351		
12	MC	0.854	-0.466	0.083		
13	MC	1.375	-0.104	0.267		
14	MC	0.776	0.397	0.067		
15	MC	0.727	-3.184	0		
16	MC	0.836	-2.966	0		
17	MC	1.024	-0.326	0.278		
18	MC	1.357	-1.095	0.129		
19	MC	0.731	0.53	0.24		
20	MC	1.042	-1.61	0.183		
21	MC	1.348	-0.009	0.146		
22	MC	1.31	-0.369	0.087		
23	MC	0.502	-1.811	0		
24	MC	0.356	-3.517	0		
25	MC	0.995	-1.061	0.141		
26	MC	1.254	-0.513	0.281		
27	MC	0.963	-1.059	0.172		
28	MC	0.68	0.382	0.225		
29	MC	1.661	0.284	0.239		
30	MC	0.526	-2.034	0		
31	MC	0.937	-0.291	0.123		
32	MC	1.011	0.704	0.197		
33	SA	0.558	-1.075			
34	SA	0.891	-0.669			
35	SA	1.07	-0.045			
36	SA	0.788	-0.344			
Item #	**Type**	**A**	**D1**	**D2**	**D3**	**D4**
37	ER	1.227	-2.001	-0.463	0.31	1.188
38	ER	1.823	-1.256	-0.453	0.176	0.592
39	ER	1.402	-0.963	-0.27	-0.115	0.368
40	ER	1.16	-2.052	-1.077	-0.003	1.013
41	ER	1.532	-1.233	-0.522	0.061	1.278
42	ER	1.287	-2.033	-0.996	-0.001	0.701

Table A.4 “True” Exemplars for Test B

Item	RP 50				RP 65				RP 80				RP 50 D				RP 65 D				RP 80 D			
#	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5	L2	L3	L4	L5
1	✓				✓				✓															
2	✓				✓					✓														
3		✓					✓				✓								✓				✓	
4	✓					✓					✓													
5		✓					✓				✓													
6			✓				✓				✓				✓				✓				✓	
7			✓				✓					✓			✓				✓					
8		✓					✓				✓								✓				✓	
9	✓				✓					✓														
10	✓				✓					✓			✓				✓							
11		✓				✓					✓													
12		✓				✓					✓			✓				✓						
13		✓				✓					✓			✓				✓						
14			✓				✓					✓												
15	✓				✓				✓															
16	✓				✓				✓															
17		✓				✓					✓													
18	✓					✓				✓			✓					✓				✓		
19			✓				✓					✓												
20	✓				✓					✓			✓				✓							
21		✓					✓				✓			✓					✓				✓	
22		✓				✓					✓			✓				✓						
23	✓				✓					✓														
24	✓				✓				✓															
25	✓					✓				✓			✓											
26		✓				✓				✓				✓				✓				✓		
27	✓					✓				✓														
28		✓					✓					✓												
29			✓				✓				✓				✓				✓				✓	
30	✓				✓					✓														
31		✓				✓					✓			✓				✓						
32			✓				✓					✓			✓				✓					
33	✓					✓					✓													
34		✓				✓					✓			✓				✓						
35		✓					✓				✓			✓					✓				✓	
36		✓					✓				✓			✓										
37_1	✓																							
37_2																								
37_3																								
37_4				✓				✓				✓				✓				✓				✓
38_1	✓												✓											
38_2																								
38_3																								
38_4			✓				✓					✓			✓				✓					
39_1																								
39_2																								
39_3																								
39_4			✓				✓				✓				✓				✓				✓	
40_1																								
40_2																								
40_3																								
40_4				✓				✓				✓				✓				✓				✓
41_1																								
41_2																								
41_3			✓																					
41_4				✓				✓				✓				✓				✓				✓
42_1																								
42_2																								
42_3																								
42_4			✓				✓					✓			✓				✓					

APPENDIX B

TABLES OF SIMULATION RESULTS

Table B.1 False Positive Results for Test A under Normal Distribution Condition

Simulated Condition		Item-mapping Criteria	Model-based Method								Empirical-based Method							
			Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
Theta Distribution	Sample Size		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Normal	1,000	RP 50	54	6	44	3	32	2	51	0	54	6	44	2	32	4	51	2
		RP 65	56	0	50	4	32	7	43	2	56	0	50	1	32	5	43	5
		RP 80	56	0	54	1	45	5	30	7	56	0	54	0	45	3	30	13
		RP 50 + Disc.	56	1	54	1	46	3	52	0	56	3	54	2	46	4	52	2
		RP 65 + Disc.	56	0	55	1	45	3	49	2	56	0	55	0	45	3	49	3
		RP 80 + Disc.	56	0	55	0	52	1	50	2	56	0	55	0	52	1	50	3
Normal	2,000	RP 50	54	6	44	2	32	1	51	0	54	5	44	1	32	2	51	3
		RP 65	56	0	50	3	32	3	43	3	56	0	50	2	32	1	43	9
		RP 80	56	0	54	0	45	2	30	5	56	0	54	0	45	2	30	14
		RP 50 + Disc.	56	2	54	1	46	1	52	0	56	1	54	0	46	1	52	4
		RP 65 + Disc.	56	0	55	0	45	1	49	1	56	0	55	0	45	1	49	3
		RP 80 + Disc.	56	0	55	0	52	0	50	1	56	0	55	0	52	0	50	3
Normal	5,000	RP 50	54	6	44	2	32	0	51	0	54	6	44	1	32	1	51	2
		RP 65	56	0	50	3	32	3	43	3	56	0	50	1	32	1	43	8
		RP 80	56	0	54	0	45	2	30	3	56	0	54	0	45	1	30	12
		RP 50 + Disc.	56	2	54	1	46	1	52	0	56	1	54	0	46	0	52	2
		RP 65 + Disc.	56	0	55	0	45	1	49	0	56	0	55	0	45	0	49	2
		RP 80 + Disc.	56	0	55	0	52	0	50	0	56	0	55	0	52	0	50	1
Normal	50,000	RP 50	54	4	44	0	32	0	51	0	54	4	44	0	32	0	51	2
		RP 65	56	0	50	0	32	0	43	2	56	0	50	2	32	0	43	12
		RP 80	56	0	54	0	45	0	30	0	56	0	54	0	45	0	30	17
		RP 50 + Disc.	56	2	54	2	46	0	52	0	56	2	54	0	46	0	52	0
		RP 65 + Disc.	56	0	55	0	45	0	49	0	56	0	55	0	45	0	49	0
		RP 80 + Disc.	56	0	55	0	52	0	50	0	56	0	55	0	52	0	50	0

N= Number of "true" non-exemplar items

% = Percentage of times incorrect items were identified

Table B.2 False Positive Results for Test A under Skewed Distribution Condition

Simulated Condition		Item-mapping Criteria	Model-based Method								Empirical-based Method							
			Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
Theta Distribution	Sample Size		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Skewed	1,000	RP 50	54	5	44	3	32	2	51	0	54	11	44	24	32	5	51	0
		RP 65	56	0	50	4	32	5	43	3	56	4	50	11	32	10	43	1
		RP 80	56	0	54	1	45	3	30	8	56	0	54	2	45	6	30	6
		RP 50 + Disc.	56	0	54	2	46	2	52	0	56	3	54	3	46	3	52	1
		RP 65 + Disc.	56	0	55	1	45	2	49	1	56	3	55	1	45	3	49	2
		RP 80 + Disc.	56	0	55	0	52	0	50	1	56	0	55	0	52	1	50	2
Skewed	2,000	RP 50	54	4	44	3	32	0	51	0	54	12	44	26	32	6	51	0
		RP 65	56	0	50	3	32	4	43	1	56	4	50	12	32	9	43	0
		RP 80	56	0	54	0	45	2	30	5	56	0	54	2	45	3	30	4
		RP 50 + Disc.	56	1	54	2	46	0	52	0	56	3	54	2	46	3	52	0
		RP 65 + Disc.	56	0	55	1	45	1	49	1	56	2	55	2	45	2	49	1
		RP 80 + Disc.	56	0	55	0	52	0	50	1	56	0	55	0	52	0	50	1
Skewed	5,000	RP 50	54	4	44	1	32	0	51	0	54	11	44	27	32	4	51	0
		RP 65	56	0	50	3	32	1	43	4	56	4	50	13	32	10	43	0
		RP 80	56	0	54	0	45	1	30	2	56	0	54	2	45	2	30	4
		RP 50 + Disc.	56	1	54	2	46	0	52	0	56	2	54	3	46	1	52	0
		RP 65 + Disc.	56	0	55	0	45	0	49	0	56	2	55	2	45	2	49	0
		RP 80 + Disc.	56	0	55	0	52	0	50	0	56	0	55	0	52	0	50	0
Skewed	50,000	RP 50	54	0	44	0	32	0	51	0	54	11	44	27	32	6	51	0
		RP 65	56	0	50	0	32	0	43	2	56	4	50	14	32	9	43	0
		RP 80	56	0	54	0	45	0	30	0	56	0	54	2	45	2	30	3
		RP 50 + Disc.	56	0	54	2	46	0	52	0	56	4	54	2	46	2	52	0
		RP 65 + Disc.	56	0	55	0	45	0	49	0	56	2	55	2	45	2	49	0
		RP 80 + Disc.	56	0	55	0	52	0	50	0	56	0	55	0	52	0	50	0

N= Number of "true" non-exemplar items

% = Percentage of times incorrect items were identified

Table B.3 True Positive Results for Test A under Normal Distribution Condition

Simulated Condition		Item-mapping Criteria	Model-based Method								Empirical-based Method							
			Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
Theta Distribution	Sample Size		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Normal	1,000	RP 50	2	95	12	71	24	95	5	90	2	100	12	69	24	93	5	92
		RP 65	0	NA	6	95	24	90	13	85	0	NA	6	85	24	88	13	93
		RP 80	0	NA	2	90	11	94	26	88	0	NA	2	85	11	81	26	91
		RP 50 + Disc.	0	NA	2	20	10	75	4	98	0	NA	2	5	10	66	4	98
		RP 65 + Disc.	0	NA	1	100	11	69	7	84	0	NA	1	80	11	61	7	91
		RP 80 + Disc.	0	NA	1	100	4	75	6	78	0	NA	1	80	4	65	6	87
Normal	2,000	RP 50	2	95	12	73	24	97	5	96	2	95	12	74	24	92	5	100
		RP 65	0	NA	6	100	24	88	13	92	0	NA	6	95	24	79	13	98
		RP 80	0	NA	2	95	11	99	26	91	0	NA	2	70	11	71	26	89
		RP 50 + Disc.	0	NA	2	5	10	74	4	100	0	NA	2	15	10	51	4	100
		RP 65 + Disc.	0	NA	1	100	11	72	7	81	0	NA	1	100	11	45	7	91
		RP 80 + Disc.	0	NA	1	100	4	88	6	70	0	NA	1	100	4	43	6	72
Normal	5,000	RP 50	2	100	12	74	24	97	5	100	2	95	12	73	24	94	5	98
		RP 65	0	NA	6	100	24	89	13	94	0	NA	6	98	24	83	13	98
		RP 80	0	NA	2	100	11	100	26	89	0	NA	2	85	11	75	26	88
		RP 50 + Disc.	0	NA	2	15	10	72	4	100	0	NA	2	0	10	47	4	100
		RP 65 + Disc.	0	NA	1	100	11	70	7	86	0	NA	1	90	11	44	7	91
		RP 80 + Disc.	0	NA	1	100	4	90	6	72	0	NA	1	90	4	43	6	75
Normal	50,000	RP 50	2	100	12	83	24	100	5	100	2	100	12	83	24	96	5	100
		RP 65	0	NA	6	100	24	96	13	100	0	NA	6	100	24	75	13	100
		RP 80	0	NA	2	100	11	100	26	92	0	NA	2	100	11	64	26	92
		RP 50 + Disc.	0	NA	2	0	10	80	4	100	0	NA	2	0	10	40	4	100
		RP 65 + Disc.	0	NA	1	100	11	82	7	100	0	NA	1	100	11	36	7	100
		RP 80 + Disc.	0	NA	1	100	4	100	6	83	0	NA	1	100	4	25	6	83

N= Number of "true" exemplar items

% = Percentage of times "true" exemplars were correctly identified

Table B.4 True Positive Results for Test A under Skewed Distribution Condition

Simulated Condition		Item-mapping Criteria	Model-based Method								Empirical-based Method							
			Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
Theta Distribution	Sample Size		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Skewed	1,000	RP 50	2	90	12	77	24	94	5	92	2	100	12	52	24	57	5	66
		RP 65	0	NA	6	98	24	87	13	88	0	NA	6	62	24	76	13	75
		RP 80	0	NA	2	90	11	95	26	91	0	NA	2	100	11	89	26	79
		RP 50 + Disc.	0	NA	2	20	10	65	4	100	0	NA	2	5	10	25	4	65
		RP 65 + Disc.	0	NA	1	100	11	63	7	71	0	NA	1	20	11	25	7	53
		RP 80 + Disc.	0	NA	1	100	4	85	6	58	0	NA	1	20	4	45	6	40
Skewed	2,000	RP 50	2	100	12	81	24	95	5	100	2	100	12	47	24	51	5	64
		RP 65	0	NA	6	100	24	91	13	91	0	NA	6	63	24	75	13	78
		RP 80	0	NA	2	100	11	99	26	90	0	NA	2	100	11	90	26	83
		RP 50 + Disc.	0	NA	2	40	10	65	4	100	0	NA	2	5	10	18	4	63
		RP 65 + Disc.	0	NA	1	100	11	60	7	76	0	NA	1	0	11	25	7	56
		RP 80 + Disc.	0	NA	1	100	4	78	6	60	0	NA	1	0	4	45	6	40
Skewed	5,000	RP 50	2	100	12	82	24	98	5	100	2	100	12	49	24	50	5	72
		RP 65	0	NA	6	100	24	87	13	97	0	NA	6	63	24	73	13	75
		RP 80	0	NA	2	100	11	100	26	87	0	NA	2	100	11	91	26	87
		RP 50 + Disc.	0	NA	2	30	10	67	4	100	0	NA	2	0	10	17	4	85
		RP 65 + Disc.	0	NA	1	100	11	64	7	87	0	NA	1	0	11	18	7	59
		RP 80 + Disc.	0	NA	1	100	4	83	6	68	0	NA	1	0	4	30	6	37
Skewed	50,000	RP 50	2	100	12	100	24	100	5	100	2	100	12	50	24	50	5	60
		RP 65	0	NA	6	100	24	96	13	100	0	NA	6	67	24	71	13	77
		RP 80	0	NA	2	100	11	100	26	88	0	NA	2	100	11	91	26	85
		RP 50 + Disc.	0	NA	2	100	10	70	4	100	0	NA	2	0	10	20	4	75
		RP 65 + Disc.	0	NA	1	100	11	73	7	100	0	NA	1	0	11	27	7	57
		RP 80 + Disc.	0	NA	1	100	4	100	6	83	0	NA	1	0	4	50	6	33

N= Number of "true" exemplar items
% = Percentage of times "true" exemplars were correctly identified

Table B.5 False Positive Results for Test B under Normal Distribution Condition

Simulated Condition		Item-mapping Criteria	Model-based Method								Empirical-based Method							
			Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
Theta Distribution	Sample Size		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Normal	1,000	RP 50	43	4	45	2	50	0	57	0	43	3	45	2	50	1	57	0
		RP 65	50	4	47	1	44	0	57	1	50	4	47	1	44	1	57	1
		RP 80	56	3	50	0	42	0	50	2	56	2	50	0	42	2	50	2
		RP 50 + Disc.	55	0	51	1	53	1	57	0	55	3	51	1	53	2	57	0
		RP 65 + Disc.	58	2	53	0	49	1	57	1	58	4	53	0	49	1	57	1
		RP 80 + Disc.	60	0	58	0	53	0	57	1	60	1	58	0	53	0	57	2
Normal	2,000	RP 50	43	3	45	3	50	1	57	0	43	2	45	4	50	1	57	0
		RP 65	50	3	47	0	44	0	57	1	50	3	47	1	44	1	57	2
		RP 80	56	3	50	0	42	0	50	1	56	3	50	0	42	1	50	2
		RP 50 + Disc.	55	0	51	2	53	1	57	0	55	1	51	1	53	0	57	0
		RP 65 + Disc.	58	1	53	1	49	1	57	1	58	3	53	0	49	0	57	1
		RP 80 + Disc.	60	1	58	0	53	0	57	1	60	1	58	0	53	0	57	1
Normal	5,000	RP 50	43	3	45	2	50	0	57	0	43	3	45	2	50	0	57	0
		RP 65	50	3	47	0	44	0	57	2	50	2	47	0	44	0	57	2
		RP 80	56	3	50	0	42	0	50	2	56	3	50	0	42	0	50	2
		RP 50 + Disc.	55	0	51	0	53	0	57	0	55	1	51	0	53	0	57	0
		RP 65 + Disc.	58	1	53	0	49	0	57	1	58	2	53	0	49	0	57	2
		RP 80 + Disc.	60	0	58	0	53	0	57	1	60	1	58	0	53	0	57	2
Normal	50,000	RP 50	43	2	45	2	50	0	57	0	43	5	45	4	50	0	57	0
		RP 65	50	2	47	0	44	0	57	2	50	2	47	0	44	0	57	2
		RP 80	56	4	50	0	42	0	50	2	56	4	50	0	42	0	50	2
		RP 50 + Disc.	55	0	51	0	53	2	57	0	55	2	51	0	53	0	57	0
		RP 65 + Disc.	58	2	53	0	49	0	57	2	58	2	53	0	49	0	57	2
		RP 80 + Disc.	60	0	58	0	53	0	57	2	60	2	58	0	53	0	57	2

N= Number of "true" non-exemplar items

% = Percentage of times incorrect items were identified

Table B.6 False Positive Results for Test B under Skewed Distribution Condition

Simulated Condition		Item-mapping Criteria	Model-based Method								Empirical-based Method							
			Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
Theta Distribution	Sample Size		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Skewed	1,000	RP 50	43	4	45	4	50	1	57	0	43	14	45	11	50	2	57	0
		RP 65	50	4	47	1	44	0	57	2	50	8	47	12	44	0	57	0
		RP 80	56	2	50	0	42	0	50	1	56	6	50	8	42	3	50	0
		RP 50 + Disc.	55	0	51	1	53	1	57	0	55	6	51	7	53	2	57	0
		RP 65 + Disc.	58	1	53	0	49	1	57	1	58	5	53	7	49	0	57	0
		RP 80 + Disc.	60	0	58	0	53	0	57	1	60	3	58	3	53	1	57	0
Skewed	2,000	RP 50	43	4	45	4	50	0	57	0	43	13	45	12	50	2	57	0
		RP 65	50	4	47	1	44	1	57	2	50	9	47	12	44	0	57	0
		RP 80	56	3	50	0	42	0	50	2	56	6	50	8	42	2	50	0
		RP 50 + Disc.	55	0	51	1	53	0	57	0	55	5	51	8	53	2	57	0
		RP 65 + Disc.	58	2	53	0	49	0	57	1	58	5	53	6	49	0	57	0
		RP 80 + Disc.	60	0	58	0	53	0	57	1	60	1	58	2	53	1	57	0
Skewed	5,000	RP 50	43	4	45	3	50	0	57	0	43	11	45	12	50	2	57	0
		RP 65	50	3	47	0	44	0	57	2	50	8	47	10	44	0	57	0
		RP 80	56	3	50	0	42	0	50	2	56	5	50	7	42	2	50	0
		RP 50 + Disc.	55	0	51	1	53	0	57	0	55	4	51	8	53	2	57	0
		RP 65 + Disc.	58	1	53	0	49	0	57	2	58	5	53	6	49	0	57	0
		RP 80 + Disc.	60	1	58	0	53	0	57	2	60	1	58	2	53	2	57	0
Skewed	50,000	RP 50	43	0	45	0	50	0	57	0	43	12	45	13	50	2	57	0
		RP 65	50	0	47	0	44	0	57	2	50	8	47	11	44	0	57	0
		RP 80	56	4	50	0	42	0	50	2	56	7	50	10	42	2	50	0
		RP 50 + Disc.	55	0	51	0	53	0	57	0	55	4	51	8	53	2	57	0
		RP 65 + Disc.	58	0	53	0	49	0	57	2	58	5	53	6	49	0	57	0
		RP 80 + Disc.	60	2	58	0	53	0	57	2	60	2	58	2	53	2	57	0

N= Number of "true" non-exemplar items

% = Percentage of times incorrect items were identified

Table B.7 True Positive Results for Test B under Normal Distribution Condition

Item-mapping Criteria	Model-based Method								Empirical-based Method							
	Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
RP 50	17	100	15	87	10	91	3	97	17	96	15	89	10	92	3	93
RP 65	10	96	13	85	16	94	3	100	10	95	13	83	16	94	3	100
RP 80	4	95	10	84	18	96	10	92	4	95	10	79	18	95	10	100
RP 50 + Disc.	5	46	9	62	7	71	3	97	5	72	9	56	7	77	3	93
RP 65 + Disc.	2	10	7	53	11	76	3	100	2	70	7	44	11	77	3	100
RP 80 + Disc.	0	NA	2	35	7	84	3	73	0	NA	2	20	7	80	3	100
RP 50	17	100	15	90	10	87	3	100	17	96	15	91	10	84	3	100
RP 65	10	100	13	89	16	94	3	100	10	98	13	84	16	91	3	100
RP 80	4	100	10	83	18	96	10	93	4	98	10	80	18	93	10	89
RP 50 + Disc.	5	48	9	74	7	67	3	100	5	62	9	67	7	60	3	100
RP 65 + Disc.	2	20	7	70	11	76	3	100	2	65	7	59	11	72	3	100
RP 80 + Disc.	0	NA	2	45	7	87	3	77	0	NA	2	25	7	81	3	63
RP 50	17	100	15	93	10	92	3	100	17	97	15	91	10	91	3	100
RP 65	10	100	13	88	16	93	3	100	10	98	13	90	16	94	3	100
RP 80	4	100	10	81	18	96	10	90	4	100	10	84	18	94	10	89
RP 50 + Disc.	5	48	9	67	7	71	3	100	5	62	9	60	7	67	3	100
RP 65 + Disc.	2	0	7	60	11	74	3	100	2	45	7	49	11	73	3	100
RP 80 + Disc.	0	NA	2	45	7	89	3	67	0	NA	2	30	7	83	3	63
RP 50	17	100	15	87	10	90	3	100	17	94	15	87	10	80	3	100
RP 65	10	100	13	92	16	94	3	100	10	100	13	92	16	94	3	100
RP 80	4	100	10	80	18	94	10	90	4	100	10	80	18	94	10	90
RP 50 + Disc.	5	40	9	67	7	71	3	100	5	80	9	56	7	57	3	100
RP 65 + Disc.	2	0	7	57	11	73	3	100	2	100	7	57	11	73	3	100
RP 80 + Disc.	0	NA	2	50	7	86	3	67	0	NA	2	50	7	86	3	67

N= Number of "true" exemplar items

% = Percentage of times "true" exemplars were correctly identified

Table B.8 True Positive Results for Test B under Skewed Distribution Condition

Item-mapping Criteria	Model-based Method								Empirical-based Method							
	Level 2		Level 3		Level 4		Level 5		Level 2		Level 3		Level 4		Level 5	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
RP 50	17	99	15	87	10	82	3	100	17	94	15	65	10	50	3	67
RP 65	10	97	13	85	16	89	3	100	10	100	13	68	16	63	3	100
RP 80	4	98	10	85	18	96	10	82	4	100	10	65	18	77	10	86
RP 50 + Disc.	5	42	9	56	7	57	3	100	5	52	9	50	7	47	3	67
RP 65 + Disc.	2	10	7	49	11	69	3	100	2	40	7	37	11	35	3	93
RP 80 + Disc.	0	NA	2	50	7	83	3	40	0	NA	2	15	7	26	3	87
RP 50	17	99	15	86	10	85	3	100	17	91	15	67	10	46	3	67
RP 65	10	97	13	82	16	93	3	100	10	100	13	66	16	66	3	100
RP 80	4	98	10	84	18	95	10	78	4	100	10	64	18	77	10	91
RP 50 + Disc.	5	44	9	52	7	60	3	100	5	46	9	49	7	44	3	67
RP 65 + Disc.	2	5	7	46	11	71	3	100	2	15	7	34	11	35	3	97
RP 80 + Disc.	0	NA	2	30	7	84	3	27	0	NA	2	0	7	24	3	93
RP 50	17	100	15	88	10	86	3	100	17	93	15	70	10	48	3	67
RP 65	10	100	13	87	16	94	3	100	10	100	13	69	16	69	3	100
RP 80	4	100	10	85	18	94	10	80	4	100	10	70	18	82	10	92
RP 50 + Disc.	5	48	9	62	7	64	3	100	5	52	9	46	7	43	3	67
RP 65 + Disc.	2	20	7	56	11	72	3	100	2	35	7	27	11	32	3	100
RP 80 + Disc.	0	NA	2	25	7	84	3	33	0	NA	2	0	7	21	3	100
RP 50	17	100	15	100	10	100	3	100	17	88	15	67	10	40	3	67
RP 65	10	100	13	100	16	94	3	100	10	100	13	69	16	69	3	100
RP 80	4	100	10	80	18	94	10	80	4	100	10	60	18	72	10	90
RP 50 + Disc.	5	80	9	78	7	71	3	100	5	40	9	44	7	43	3	67
RP 65 + Disc.	2	50	7	86	11	73	3	100	2	0	7	29	11	27	3	100
RP 80 + Disc.	0	NA	2	100	7	86	3	33	0	NA	2	0	7	14	3	100

N= Number of "true" exemplar items

% = Percentage of times "true" exemplars were correctly identified

APPENDIX C

SAMPLE AA, BB, AND CC PLOTS

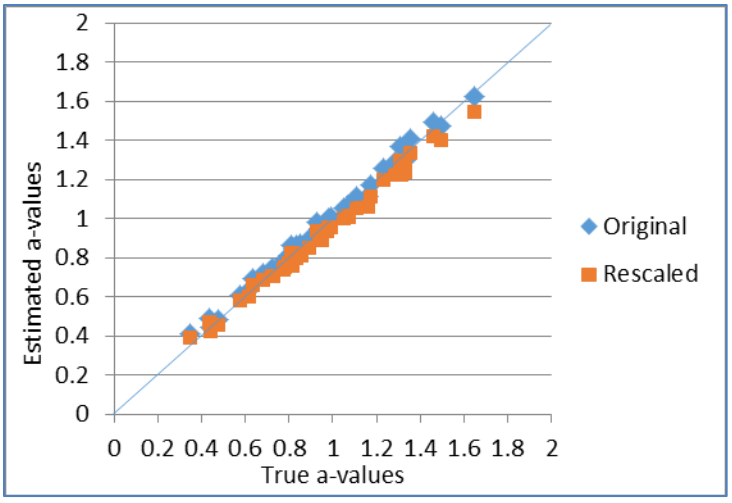

Figure C.1 aa Plot for Test A under Normal, 50K condition

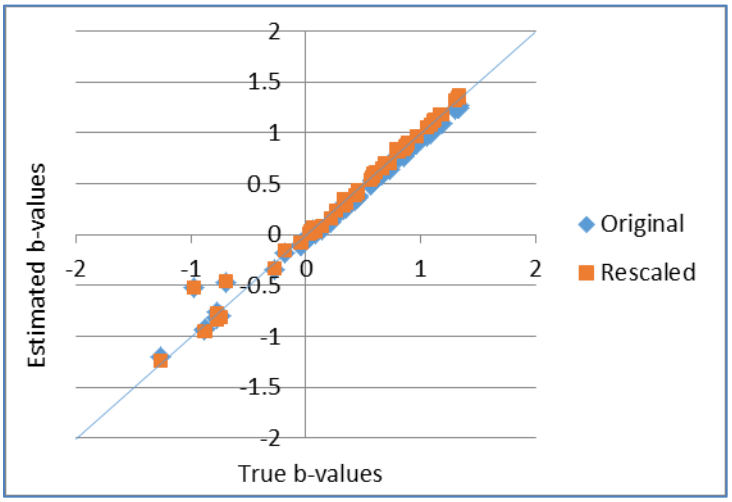

Figure C.2 bb Plot for Test A under Normal, 50K condition

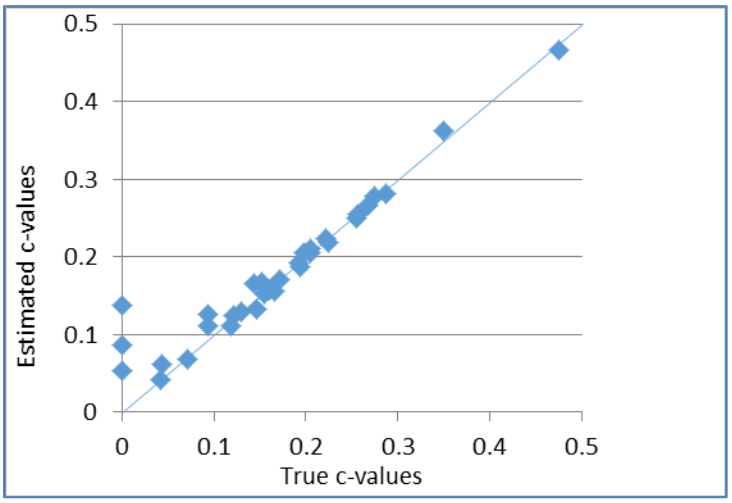

Figure C.3 cc Plot for Test A under Normal, 50K condition

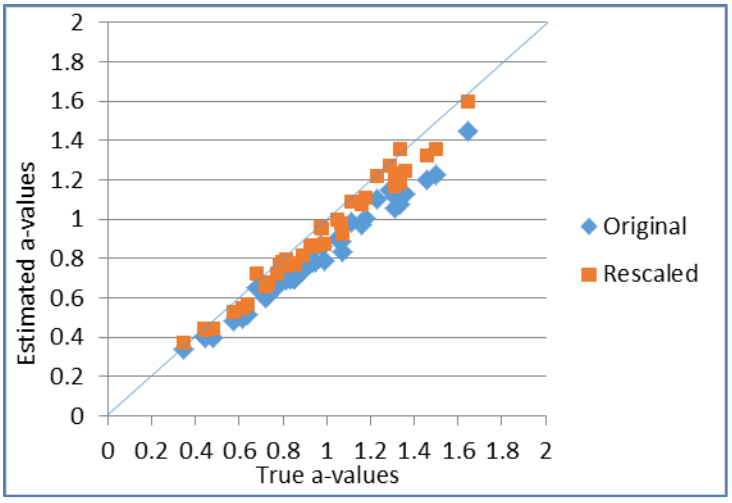

Figure C.4 aa Plot for Test A under Skewed, 50K condition

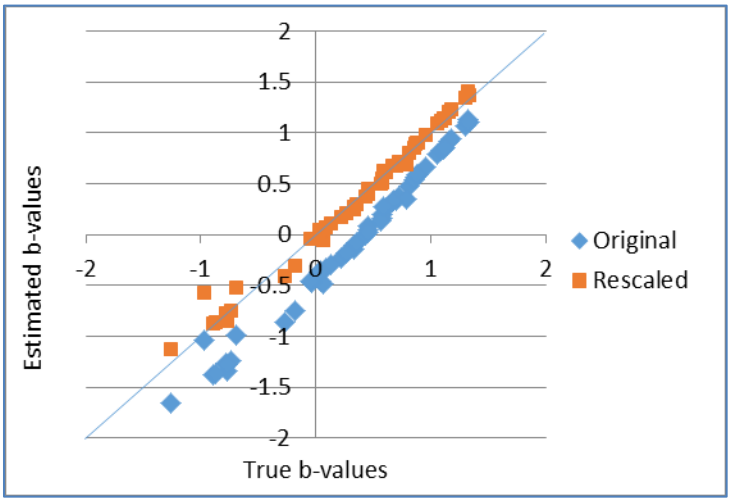

Figure C.5 bb Plot for Test A under Skewed, 50K condition

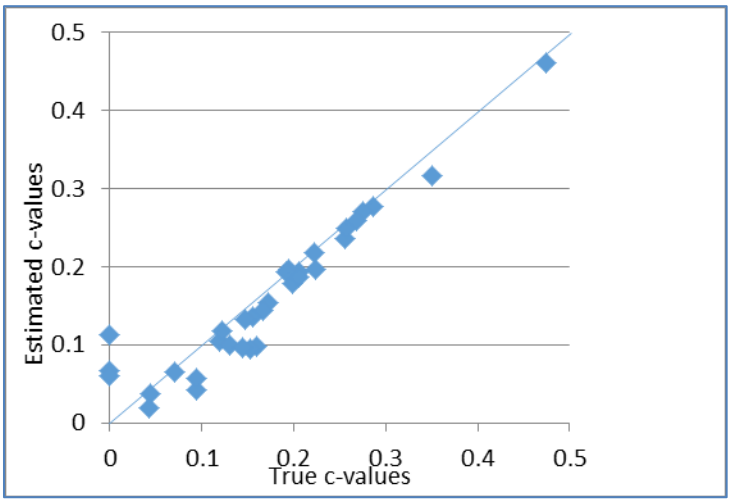

Figure C.6 cc Plot for Test A under Skewed, 50K condition

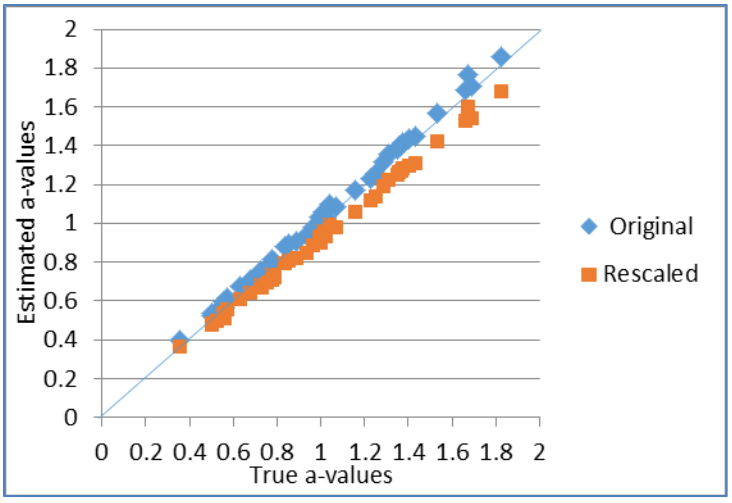

Figure C.7 aa Plot for Test B under Normal, 50K condition

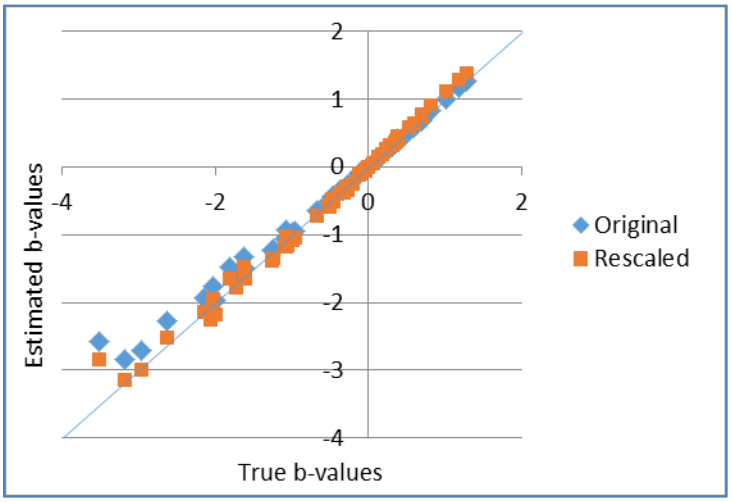

Figure C.8 bb Plot for Test B under Normal, 50K condition

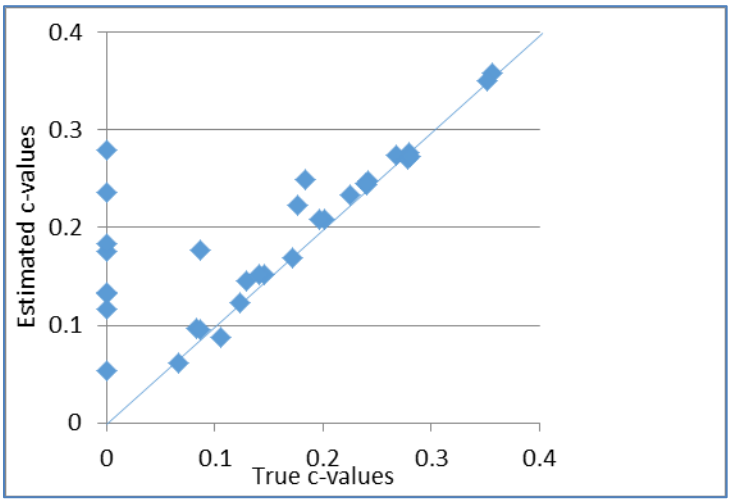

Figure C.9 cc Plot for Test B under Normal, 50K condition

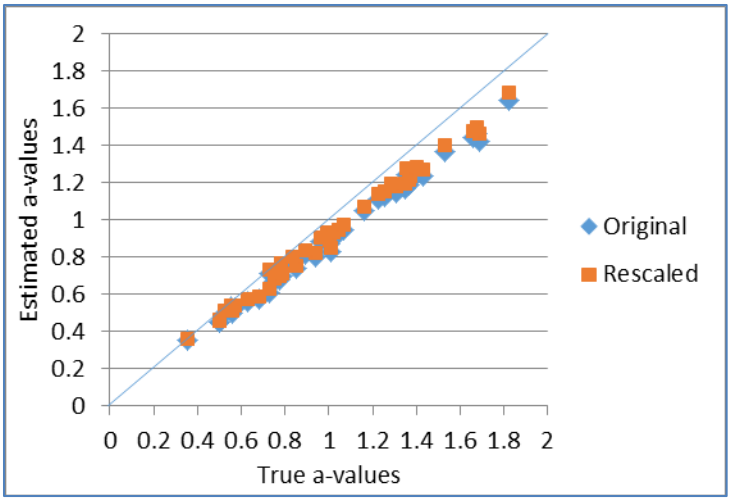

Figure C.10 aa Plot for Test B under Skewed, 50K condition

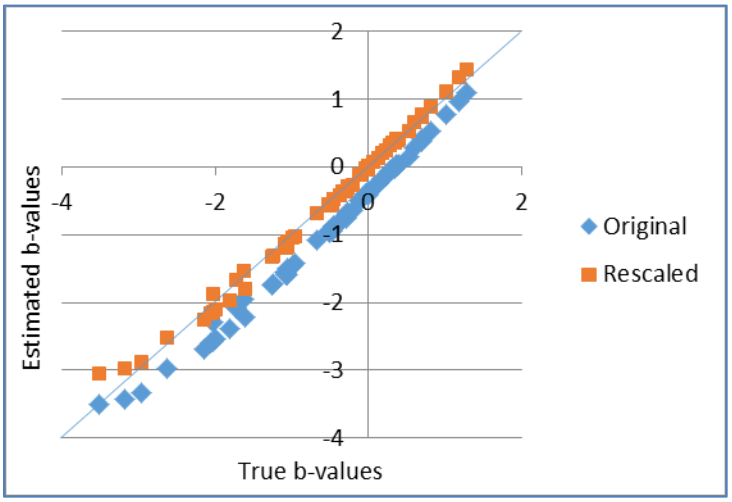

Figure C.11 bb Plot for Test B under Skewed, 50K condition

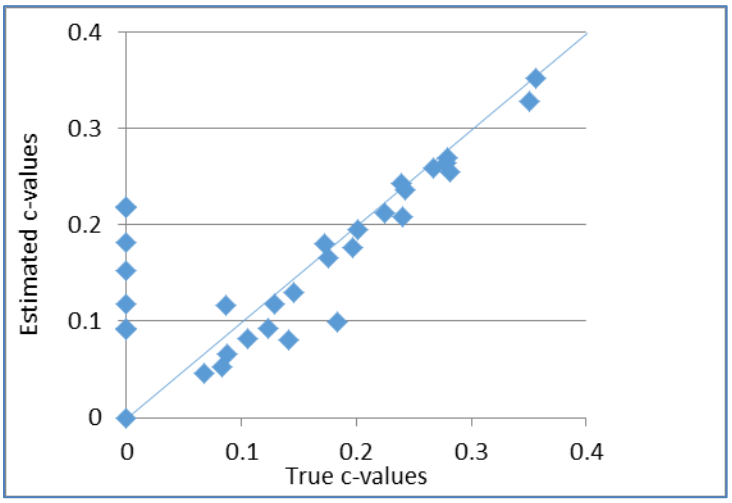

Figure C.12 cc Plot for Test B under Skewed, 50K condition

BIBLIOGRAPHY

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191-204.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement, 29*(2), 163-175.

Beretvas, N. S. (2004). Comparison of Bookmark difficulty locations under different item response models. *Applied Psychological Measurement, 28*(1), 25-47.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., & Harris, E. L. (1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3*(1), 9-51.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355-386). Westport, CT: Praeger Publishers.

Every Student Succeeds Act of 2015, Public Law No. 114-95 (2015).

Forte Fast, E., Blank, R. K., Potts, A., & Williams, A. (2002). *A guide to effective accountability reporting*. Washington, DC: Council of Chief State School Officers.

Forsyth, R. A. (1976). *Describing what Johnny can do.* Iowa City, IA: Iowa Testing Programs.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice, 10*(3), 3-9, 16.

Garcia Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417–444.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Erlbaum.

Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scales and reports more understandable? In R. W. Lissitz & W.D. Schafer (Eds.), *Assessment in educational reform* (pp. 192-205). Boston: Allyn & Bacon.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., & Zwick, R. (2000). A response to "setting reasonable and useful standards in the National Academy of Sciences' "Grading the Nation's Report Card." *Educational Measurement: Issues and Practice, 19*(2), 5-14.

Hambleton, R. K., Sireci, S., & Huff, K. (2008). *Development and validation of enhanced SAT score scales using item mapping and performance category descriptions* (Final Report). Amherst: University of Massachusetts, Center for Educational Assessment.

Hambleton, R. K., & Slater, S. (1994). *Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved?* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A researched-based approach to score report design. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology.* Washington, DC: APA.

Han, K. T. (2006). WinGen2: Windows software that generates IRT parameters and item responses [computer program]. Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment. Retrieved May 13, 2007, from http://www.umass.edu/remp/software/wingen/

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23*(1), 35-56.

Huynh, H. (2000, April). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard settings*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark Standard Setting Method: A Literature Review. *Educational Measurement: Issues and Practice, 25*(1), 4-12.

Kolstad, A. (1996, April). *The response probability convention embedded in reporting prose literacy levels from 1992 National Adult Literacy Survey.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: American Institutes for Research.

Koretz, D. (1995). The quality of information from NAEP: Two examples of work done in collaboration with Leigh Burstein. *Educational Evaluation and Policy Analysis, 17*(3), 280-294.

Koretz, D., & Diebert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991.* Santa Monica, CA: RAND.

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education, 11*(1), 23-47.

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 681-699). Westport, CT: Praeger Publishers.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131-154.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Mueller, C. D., Schneider, M. C., & Egan, K. (2008, March). *Response probability criterion and subgroup performance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City.

Mullis, I. V. S., Beaton, A. E., Goodison, J. M., Johnson, E. G., MacDonald, W. B., & Mislevy, R. J. (1990). *The NAEP guide: A description of the content and methods of the 1990 and 1992 assessments.* Princeton, NJ: Educational Testing Service.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1* [computer software]. Chicago, IL: Scientific Software International.

National Academy of Education. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education Panel on the evaluation of the trial state assessment: An evaluation of the 1992 achievement levels*. Palo Alto, CA: Stanford University, The National Academy Press.

National Education Goals Panel. (1998). *Talking about tests: An idea book for state leaders*. Washington, DC: U.S. Government Printing Office.

National Research Council. (2001). *NAEP reporting practices: Investigating district-level and market –basket reporting*. Washington, DC: National Academy Press.

National Research Council. (2005). *Measuring literacy: Performance levels for adults.* Committee on Performance Levels for Adult Literacy, R.M. Hauser, C.F. Edley, Jr., J.A. Koenig, & S.W. Elliott (Eds.). Washington, DC: The National Academies Press

No Child Left Behind Act of 2001, Public Law No. 107-110 (2002).

Patelis, T. & Matos-Elefonte, H. (April, 2009). Efforts to Produce Relevant Score Reports to School, District, and State Officials on National Tests. Presentation at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Philips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Pommerich, M., Nicewander, W. A., & Hanson, B. A. (1999). Estimating average domain scores. *Journal of Educational Measurement, 36*(3), 199-216.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies: Final Report*. Greensboro, NC: SERVE.

Ryan, J. M. (2006). Practices, issues, trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 677-710). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*(4), 347-362.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science*, 5(3), 327-339.

Wainer, H. (1990). Graphical visions from William Playfair to John Tukey. *Statistical Science, 5*(3), 340-346.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14-23.

Wainer, H. (1997a). Improving tabular displays, with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*(1), 1-30.

Wainer, H. (1997b). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* New York, NY: Copernicus Books.

Wainer, H., Hambleton, R.K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*(4), 301-335.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*(3), 231-253.

Williams, N. J., & Schulz, E. M. (2005, April). *An investigation of response probability (RP) values used in standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 31,* 21-26.

Zenisky, A. L., & Hambleton, R. K. (2013). From "Here's the Story" to "You're in Charge": Development and maintaining large-scale online test and score reporting resources. (pp. 175-185). In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment: Theory, issues, and practice*. London: Taylor & Francis.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22 (4),* 359–375.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*(2), 205-218.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15-25.